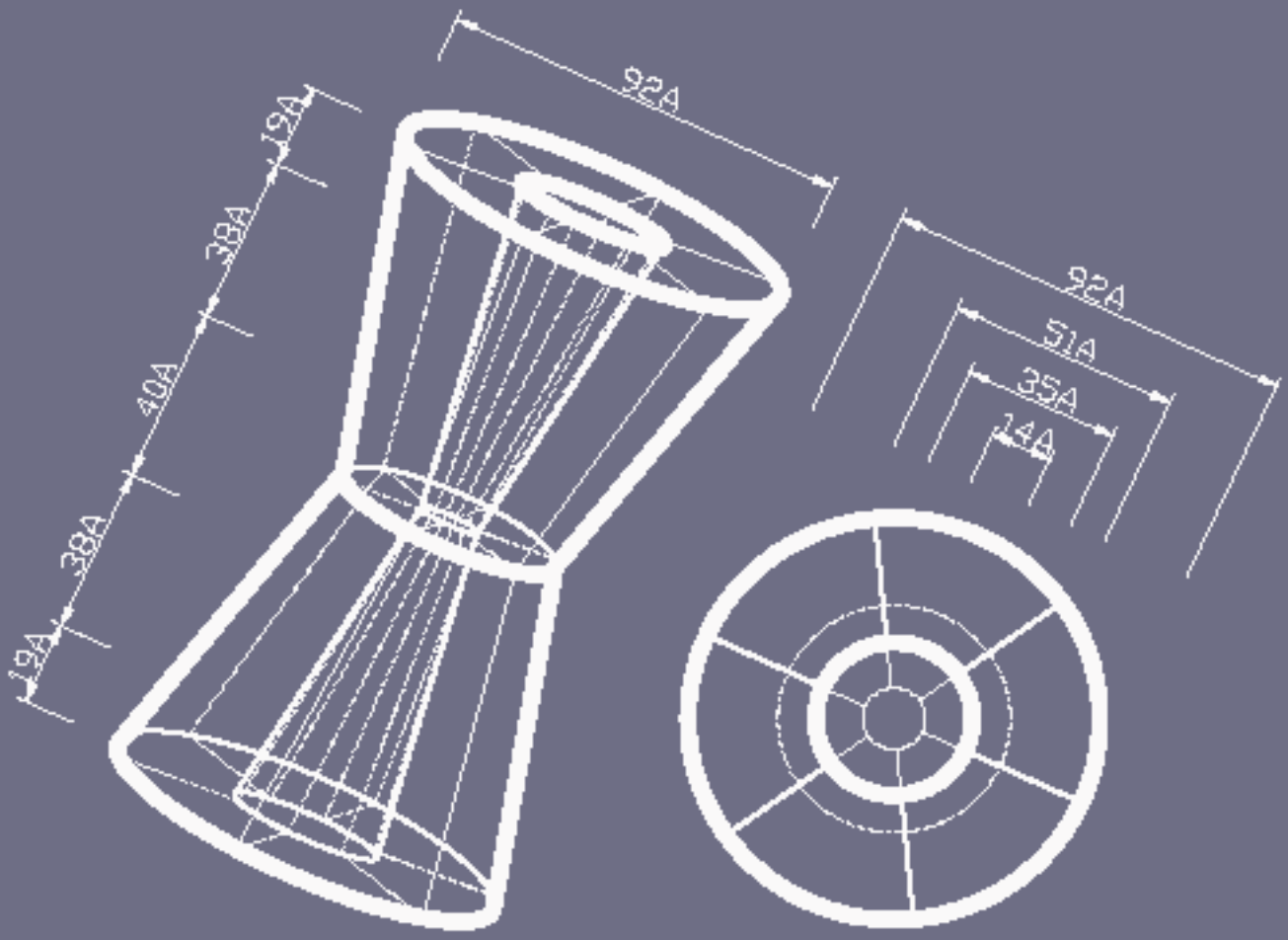
Paulina Pluta



COMPUTATIONAL AND BIOPHYSICAL METHODS IN CHARACTERIZATION OF PROTEIN-PROTEIN AND PROTEIN-LIGAND INTERACTIONS

PhD thesis
2015

Universidad del PaisVasco * Facultad de Ciencia y Technologia * Departamento de Bioquimica y Biologia Molecular

Cover design: Paulina Pluta & Michał Pluta

Thesis Doctoral

Métodos computacionales y biofísicos para la caracterización de interacciones proteína-proteína y proteína-ligando

Memoria presentada por

Paulina Pluta

Dirigida por

Dr. Óscar Millet Aguiler-Galindo

CIC bioGUNE

Para optar al titulo de Doctora por la

Facultad de Ciencia y Tecnología

Departamento de Bioquímica y Biología Molecular

Programa de Doctorado en Biología Molecular y Biomedicina

Leioa, Diciembre de 2015

Resumen de la tesis

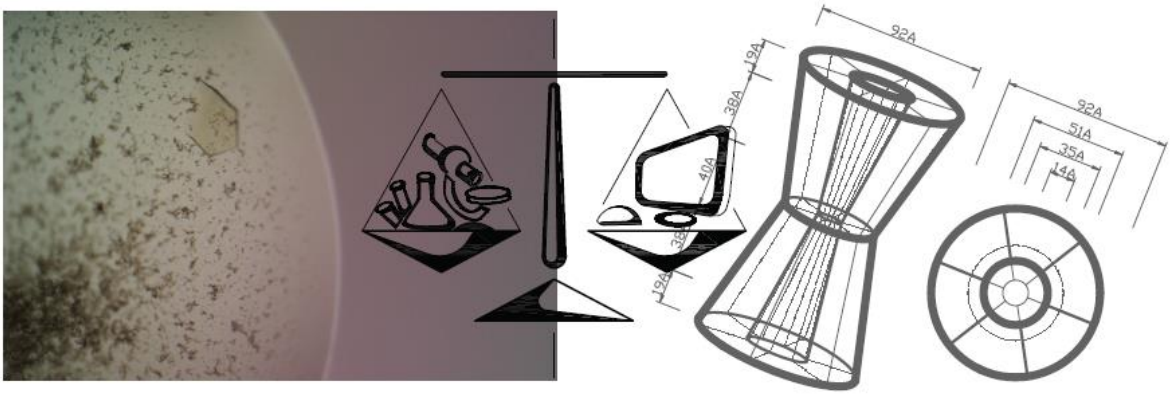


Figura 1. El equilibrio entre metodos computacionales y biofísicos en la caracterización de las interacciones moleculares como la premisa básica de la tesis; version ilustrada subjetiva de la autora.

Todas las moléculas - tan simples como la molécula de hidrógeno o tan extremadamente complejas como una proteína - resultan de la unión de átomos individuales. Los átomos están unidos mediante fuerzas intramoleculares (enlaces iónicos, metálicos o covalentes, principalmente). Estas son las fuerzas que se deben vencer para que se produzca un cambio químico, es decir, la transformación de unas moléculas en otras. Los enlaeas covalentes son muy estables desde el punto de vista energético: se requiere mucha energía para romperlos. Sin embargo, existen tambien otras fuerzas intermoleculares, más débiles, que actúan sobre distintas moléculas o iones y que hacen que éstos se atraigan o se repelan. Estas son las fuerzas que determinan en gran medida las propiedades físicas de las sustancias biológicas. Son responsables de la estructura final de macromoléculas, de la unión específica entre moléculas, es decir, del reconocimiento molecular; de los procesos de autoorganización y ensamblaje de estructuras macromoleculares y celulares y de todo lo que suponga movimiento y comunicación a nivel celular y subcelular. Su importancia, por tanto, es fundamental para todos los aspectos estructurales y funcionales de los seres vivos. Por lo general son fuerzas de carácter débil pero, al ser muy numerosas, su contribución cooperativa es muy significativa y determinante.[1,2]

Las proteínas son macromoléculas biológicas formadas por unidades denominadas aminoácidos y juegan un papel esencial en el funcionamiento de cualquier organismo. Adoptan una estructura espacial tridimensional organizada y ordenada, estrechamente relacionada con su función específica. La función desempeñada por cada proteína puede estar regulada mediante la interacción con otras proteínas y/o moléculas de bajo peso molecular (moléculas orgánicas pequeñas o metales) denominadas ligandos, cuya función puede ser de regulación (actuando como activadores e inhibidores), constitutiva (cofactores), u otra, de modo que dicha interacción regula la función de

[1] http://www.ehu.eus/biomoleculas/moleculas/fuerzas.htm
[2] http://bioquibi.webs.ull.es/temascompletos/InteraccionesNC/inicio.htm

la proteína. La interacción con otras moléculas puede conllevar un cambio en la conformación tridimensional de una proteína y, por tanto, modula su función. El mal funcionamiento de una o varias proteínas derivado de cambios en la composición de aminoácidos (mutación) es la causa de numerosas enfermedades genéticas. Las mutaciones pueden provocar un plegamiento erróneo de la molécula o una alteración estructural que afecta a la función, desencadenando un proceso patológico. Entender el comportamiento de las proteínas y sus interacciones abre una puerta hacia el desarrollo racional de tratamientos adecuados: 1) introducciendo en el organismo una versión correcta de la proteína (terapia génica); 2) inactivando una proteína (inhibición), 3) rescatando la conformación estructural y la función correcta mediante fármacos.[3]

Existe una gran variedad de métodos para descubir, analizar y caracterizar las interacciones intra- u intermoleculares dependiendo de los propósitos que nos motivan. La caracterización experimental de biomoléculas es posible con resolución atómica, mediante el uso de técnicas tales como la cristalografía de rayos X o la RMN, o a menor resolución, técnicas cinéticas (velocidad de reacción) o termodinámicas (equilibrio químico). Por otro lado, para el estudio teórico de interacciones intermoleculares y la caracterización y el análisis de la energética de unión se usan a menudo los métodos computacionales. De todos modos, los mejores resultados siempre se obtienen de la combinacion razonable de los dos métodos teoreticos y experimentales.

Mis estudios doctorales en el proyecto ***"Métodos computacionales y biofísicos para la caracterización de interacciones proteína-proteína y proteína-ligando"*** se centran en la busca del equilibrio entre dichos métodos. Concretamente me he centrado en la investigación de las interaciones entre proteinas (en el ejemplo de las conexinas) y las interaciones entre una proteina y su ligando (en el ejemplo de las enzimas involucradas en la biosíntesis del grupo hemo).

CONEXINAS

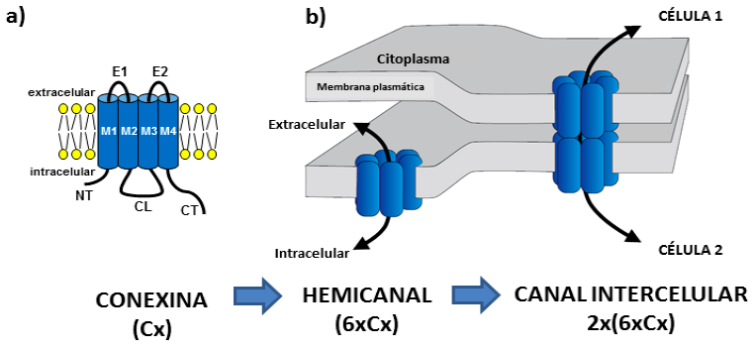


Figura 2. **Canales de conexina. Canales de conexina.** a) Todas las conexinas comparten la misma topología transmembranal, consistente en 4 dominios transmembrana (M1-M4) , un asa citoplasmática (CL), dos asas extracelulares (E1, E2) y el dominio del amino- y del carboxilo-terminal orientados hacia el interior celular. b) Los canales intercelulares se forman en dos etapas; primero se forman los precursores o hemicanales por la oligomerización de seis subunidades de conexina en cada una de las células, y luego el canal intercelular al unirse los dos hemicanales. Los hemicanales de algunas conexinas pueden también ser funcionales formando una vía de señalización entre el interior y el exterior celular.[4]

[3] http://oldwww.bifi.es/research/protein_ligand_inter/protein_ligand_inter_es.php

[4] Conectado por conexinas: "Comunicación intercellular mediada por canales de conexinas y enfermedades asociadas a mutaciones en los genes de conexina", Luis C. Barrio y Daniel González-Nieto, Unidad de Neurología Experimental, Hospital "Ramón y Cajal", Articulo divulgado en el Mayo 2012, non-indexed journal

Las conexinas forman canales intercelulares que, a diferencia de otros tipos de canales iónicos, atraviesan dos membranas plasmáticas y posibilitan la comunicación directa los citoplasmas de las células adyacentes (Figura 2). Por el poro de estos canales pueden difundir iones, metabolitos, segundos mensajeros y en general moléculas de hasta 1kDa. Este tipo de comunicación directa de *célula a célula* está presente en la gran mayoría de los tejidos y tipos celulares y participa en multitud procesos, entre los que se incluye el control de la proliferación y la diferenciación celular, la transmisión de señales eléctricas, la coordinación de la actividad metabólica y el mantenimiento de la homeostasis celular. La relevancia que actualmente se le reconoce a esta familia de proteínas deriva, en parte, del descubrimiento de varias patologías causadas por mutaciones en distintos genes de conexina. Entre estas enfermedades cabe destacar las que cursan con defectos de la mielina, pérdida de la audición, alteraciones dermatológicas, cataratas y otros síndromes mas complejos que pueden afectar a varios tejidos u órganos. Comprender el amplio rango de los defectos genéticos y la diversidad de los efectos de las mutaciones es importante para los pacientes y el consejo genético, para entender la patogénesis de la enfermedad y, en última instancia, para el desarrollo de terapias especificas.[4]

En colaboración con el grupo experimental de Dr. Luis Barrio y usando métodos computacionales hemos buscado las relaciones entre las mutaciones, la estructura y el funcionamiento. Debido al elevado peso molecular y la localizacion transmembranal de la biomolécula, los canals resultan casi imposibles de crystalizar; hasta ahora solo una structura ha sido depositada – Cx26 con el PDB ID 2ZW3[5]. Con esta structura como una base, hemos creado una serie de modelos para otras conexinas y canals (Cx32, Cx47). Para predecir y describir la energetica de unión y sus cambios causados por diferentes mutaciones hemos empleado algoritmos computacionales de "molecular docking", basados en diferentes campos de fuerza. De esta forma es posible validar diferentes modelos teóricos utilizando los datos obtenidos por la caracterización experimental.

Los objetivos del presente proyecto son:

- Caracterizar las interacciones proteína-proteína entre hemicanales creados por las proteinas de diferente masa molecular (Cx26, Cx32 y Cx47) usando “molecular docking” procedure (en comparación con los resultados experimentales).
- Caracterizar el efecto de la mutación sobre el proceso de acoplamiento de los hemicanales.
- Caracterizar la función de los diferentes motivos estructurales de las cadenas de conexina, utilizando “molecular docking” y simulaciones de dinámica molecular.

Los resultados obtenidos son:

- Usando “molecular modelling” hemos creado modelos structurales homólogos para las conexinas Cx26, Cx32 y Cx47. Hemos establecido un protocolo para generar todos los tipos de canales (homo-/heteromerico, homo-/heterotípico) que posteriormente han servido para la caracterizacion del proceso acoplamiento de los hemicanales.
- Se encontró que la energía puede ser considerada como un descriptor de la fuerza de acoplamiento dado que refleja la complementariedad energética entre los dos hemicanales

[5] Song G, Li Y, Cheng C, Zhao Y, Gao A, Zhang R, Joachimiak A, Shaw N & Liu Z-J (2009) Structural insight into acute intermittent porphyria. FASEB J 23, 396–404.

en el canal. La bondad del ajuste para este parámetro y los rendimientos experimentales para la formación GJ revelan que este es un factor clave para la formación de los canales, también explicando los aspectos importantes de la "promiscuidad" de las conexina. Por otra parte, tal acuerdo valida el novedoso método que se ha utilizado.

- Mis estudios confirman la relevancia de los giros extracellulares en la formación de canales y el papel crucial del residuo Asn176 de Cx26 (y los correspondientes en otras isoformas de Cx) y residuos Asn54 de Cx26 para la estabilidad del canal. También revelan que la composición hemicanal sólo es parcialmente responsable de la formación de canal, mientras que la disposición relativa de conexinas en los hemichannels en última instancia definen la compatibilidad hemicanal.
- Compatible con los estudios funcionales, el análisis de las trayectorias de MD para la proteína hemicanálica Cx47 nativa y sus versiones alteradas demostró la importancia del dominio N-terminal y del primer dominio transmembrana para el buen funcionamiento de los hemicanales. Las simulaciones que se han llevado a cabo identifican, por primera vez para la proteína Cx47, los residuos cruciales para la organización del poro.

ENZIMAS DE LA BIOSÍNTESIS DEL GRUPO HEMO

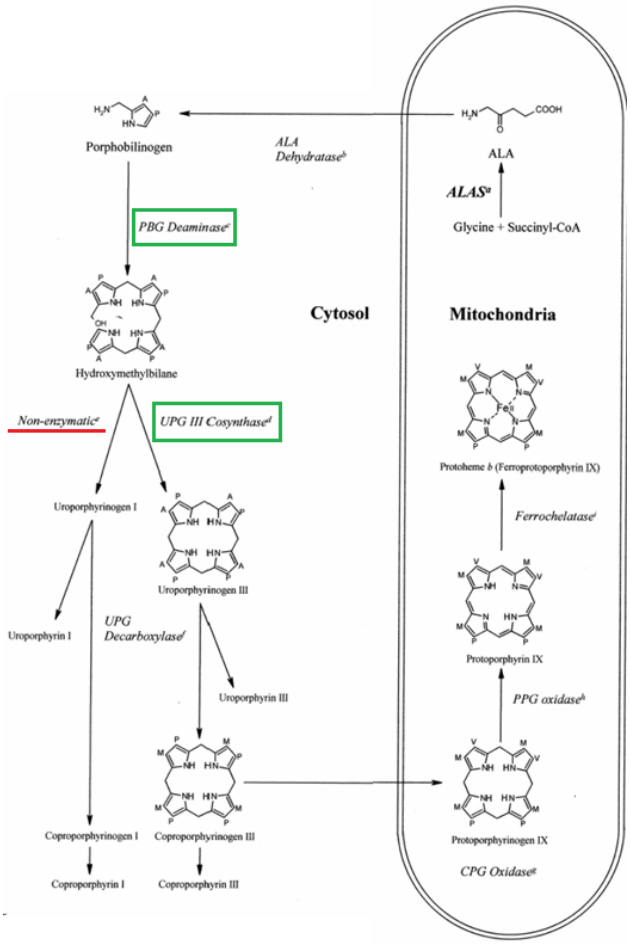


Figura 3. Resumen de la biosíntesis del grupo hemo.[6] La deaminasa porfobilinógeno (PBGD) y la uroporfirinógeno sintasa (UROIIIS) marcadas en verde. Paso patógeno marcado en rojo.

[6] Autor de la ilustracion desconocido.

El grupo hemo y sus derivados son imprescindibles para la vida y el proceso que lleva a su producción se encuentra altamente conservado en todos los seres vivos (Figura 3). El 85% del grupo hemo se sintetiza en células eritroides y se destina para la producción de hemoglobina, mientras que el 15% restante sirve como grupo prostético de enzimas dependientes del citocromo P450 y de otras hemoproteinas. La pérdida (parcial) de la actividad enzimática de alguno de las ocho enzimas involucrados conlleva la aparición de una patología denominada genéricamente porfíria. Dicha deficiencia funcional está asociada a una mutación en la secuencia de aminoácidos del enzima y la patología específica varía en función del enzima que contiene la mutación. Las porfírias están catalogadas como enfermedades raras, graves y generalmente incurables. Todos los pacientes de porfíria adolecen de anemia hemolítica y fotosensibilización de la piel. Sin embargo, la sintomatología completa depende de la variante específica de porfíria.[7,8]

Mi trabajo de tesis se centra en una de las líneas principales de investigación del grupo, el estudio de dos de las enzimas involucradas en la biosíntesis del grupo hemo - la porfobilinógeno deaminasa (PBGD) y la uroporfirinógeno sintasa (UROIIIS) causantes, respectivamente, de la porfiria aguda intermitente y la porfiria eritropoyética congénita - desde el punto de vista sintomatológico, la porfíria más severa. El objetivo es entender el mecanismo de la ruta enzimática para así poder modularla. Para esto hemos utilizado ensayos enzimáticos para medir la actividad de las enzimas *in vitro* e *in vivo*, cristalografía de rayos X para obtener información estructural de alta resolución sobre los distintos enzimas y métodos computacionales para el diseño y caracterización de inhibidores.

Los objetivos concretos del proyecto son:

- Diseño, optimización y validado de inhibidores de la PBGD.
- Determinación y validado de los residuos claves para la catálisis de la PBGD.
- Separación y caracterización de los complejos intermedios de la reacción catalizada por la PBGD.
- Caracterización de la estabilidad de los mutantes más común de la UROIIIS que producen porfíria eritropoyética congénita utilizando métodos computacionales.

Los resultados obtenidos son:

- A partir del análisis de la interacción molecular, se ha identificado una lista de compuestos que exhiben potencial inhibición de la PBGD por analogía de sustrato. También se ha detectado y caracterizado un sitio de unión secundario, activo en la acción inhibitoria.
- Los ensayos enzimáticos con el compuesto seleccionado indican inhibición competitiva y, por tanto, confirma la interacción observada *in silico*. Por otra parte, estos resultados indican que la PBGD se puede inhibir eficazmente por moléculas orgánicas pequeñas; compuestos ensayados constituyen una nueva línea de inhibidores de la cual 2,5-dimetil-1H-pirrol que demuestran el mayor nivel de inhibición en la relación a su concentración en la solución de reacción se ha establecido como una referencia para ensayos adicionales.

[7] Desnick, R. J.; Astrin, K. H. Congenital erythropoietic porphyria: advances in pathogenesis and treatment. Br J Haematol 2002, 117, 779-95.
[8] Warner, C. A.; Yoo, H. W.; Roberts, A. G.; Desnick, R. J. Congenital erythropoietic porphyria: identification and expression of exonic mutations in the uroporphyrinogen III synthase gene. J Clin Invest 1992, 89, 693-700.

- Las modificaciones del protocolo para la purificación de la PBGD han permitido la separación y el análisis de los intermediatos de la reaccion catalizada por el enzima, aptos para los estudios estructurales. Los resultados demuestran el carácter dinámico del enzima y representan – por primera vez – datos de difracción de los intermedios de la reacción.
- Estudios de la UROIIIS permitieron comprender la delicada interacción entre Cys73 y la estabilidad cinética de la proteína y demostraron que las interacciones electrostáticas no covalentes en este sitio son capaces de restituirr parcialmente la estabilidad y la actividad de la enzima.

Thesis Doctoral

Computational and biophysical methods in characterization of protein-protein and protein-ligand interactions

Paulina Pluta

Supervised by:

Dr. Óscar Millet Aguiler-Galindo

CIC bioGUNE

Leioa, Diciembre de 2015

Life is about interactions….

Summary

The thesis, "Computational and biophysical methods in characterization of protein-protein and protein-ligand interactions", involves two projects. The first one, run in collaboration with Dr. Luis Barrio (Ramón y Cajal University Hospital, Madrid), aims to explain – using experimental and computational data – the basic interactions between connexins in ion channels and the influences of pathogenic mutations discovered in clinical cases. The second project intends to study porphobilinogen deaminase (PBGD) and uroporphyrinogen III synthase (UROIIIS), the enzymes of the heme biosynthesis pathway, focusing on the structural and functional analysis with an emphasis on the catalytic mechanism and interaction with potential inhibitors.

THE CONNEXIN PROJECT. Connexins are generally considered to mediate cell-to-cell communication (critical for the maintenance of tissue/organ homeostasis) by forming gap junction channels. The assembly of connexins into channels occurs in two main stages: connexins first oligomerize into hexameric hemichannels prior to reaching the cell surface, then two hemichannels from neighbouring cells dock to generate a complete intercellular channel. The connexins belong to a multigene family with more than twenty classes identified. Although they share similar topology, different types of connexins form different types of channels with ultimately different biophysical properties of conductance, gating or permeability. This complexity gives functional diversity and cellular relevance but it also results in difficulties in conducting experiments; till now there is only one structure successfully crystallized and refined (Maeda *et al.*).

The theoretical experiment I have performed is based on this structure and the experimental data provided by Dr. Barrio's group. They reproduced different channels in cells, assaying with mutations that are suspected to have strong influence on the structure and/or to be related to clinical cases. It gave them information on which hemichannels are able or not to form functional channels. Based on the structure of Cx26 and using homology modelling (CPH and SWISS-MODEL servers), I have created a set of structural models for connexins 26, 32 and 47 (respectively, Cx26, Cx32, Cx47 – the number goes to the weight in kDa) and their mutants. Each mutation has been progressively introduced to different connexin chains in different configurations in hemichannels and docked (ClusPro Server), simulating the gap junction interactions – some channels did not dock (according to the experiments) while the others succeeded, even with the same mutation. To understand where these differences come from and if there is any pattern I have compared the docking interaction energies, contact potential and the nets of hydrogen bonds appearing in each case, both in relation to the wild type channel. The analysis of the protein–protein docking has given very good correlation with the experimental data of Dr. Barrio and has let us propose a model of interaction and bringing us closer to a better understanding of their structure-function relationships (a manuscript in preparation).

THE PBGD PROJECT. Two out of eight enzymes of the heme pathway – PBGD and UROS – have been objects of interest of our group for years (Fortian *et al.*; Blouin *et al.*). Mutations in their structures cause porphyrias, a group of rare diseases. Till now they are largely incurable and their symptoms (neurological complications and/or skin problems) hamper much of the normal life of patients. The aims of our research are: 1) to find and characterize compounds which may be

inhibitors or chaperons (in case of UROS) of the proteins and thus contribute towards therapy development, 2) to investigate structures of the enzymes, shedding light on the mechanisms of reaction and inhibition. To achieve the goals we have combined different methods and established collaborations; of which I am responsible for UROS MD simulations (ben Bdira *et al.*) and all the modelling and crystallography work of PBGD.

The most interesting characteristic of PBGD and the one causing most problems in the investigation is the reaction it catalyses. PBGD forms a linear tetrapyrrole from four units of porphobilinogen using a single active site and repeating the same action four times. The existence of the intermediates were proven before (Jordan *et al.*) but never at the structural level. Trying to fill this gap, we have established and optimized protocols for expressing and purifying wild type and mutant PBGDs, obtaining abundant yield and separating the intermediates of the reaction. It has allowed us to prepare pure and highly concentrated samples, to be used later on for activity assays, crystallization, SAXS and NMR experiments. The results we have got prove the dynamic character of the enzyme and agree with the previous predictions. The quality of data (in case of NMR) and the resolution (in case of SAXS) were not enough to explore structural details however, in collaboration with Dr. Adriana Rojas from the CIC bioGUNE crystallography platform and Prof. Richard Pickersgill's group (Queen Mary University of London) we have recently obtain crystals for some reaction intermediates which models are being refined now.

Aside from structural studies, in our laboratory we have been conducting massive screenings of structures that may potentially interact with the enzymes. Both, molecular docking and assaying ligand libraries have been undertaken. Further in vivo tests, as well as crystallization and simulations trials for the most promising candidates are planned.

INTRODUCTION

1

In modern biology, research is an indirect, phenomenological science: it measures the response of a stimulus to an organism, a cell or a molecule and the consequence is recorded using a technical measurement readout[1]. Incorporating the principles of biochemistry and biophysics, structural biology investigates properties of macromolecules, particularly proteins and nucleic acids. Diversity and complexity of life contradicts the fact that it is structurally based on a relatively small amount of basic blocks – e.g. interactions between only 20 amino acids determine structure of a protein. Same, its function and dynamics. The foundation of structural biology is reductionism; it operates on isolated systems and reduced facts with the intention to build a landscape representation, a model of molecular phenomena[2].
Over time, a whole range of methods have been invented to explore and analyse the biomolecular synergy. Biophysical methods help us to understand and characterize how the collections of inanimate particles interact to maintain life animated solely by the chemical and physical laws[3]. Ultimately, technical and theoretical developments are transforming the field, which is rapidly progressing from a science performed almost entirely at the laboratory bench to one that may be explored, at least partially, through computers[4]. Computational methods – maintaining the reductionist approach – look for patterns and algorithms extrapolating practical results where laboratory techniques are forceless or too expensive. Throughout decades the advances in computation power have resulted in more accurate methods; with current computational methods being able to look deeper, work faster and to describe in detail some mechanisms by which healthy cells function and diseases progress. Anyhow, the most reliable and accurate results are always obtained by combining both, biophysical and computational methods.

1.1 Biophysical and computational methods: main assumptions

In science it is always useful to take the view that everything can be treated as a system. A system is a set of connected elements that form a complex group. Its properties (that can be measured) are observables are the ground base used to generate models (partial descriptions of systems). A good model is an abstraction at an appropriate level of detail that accurately represents the reality of the system of interest. Empirical testing will validate or falsify the models by comparing predictions derived from theory (hypotheses) to the measured experimental data.[2]
From the practical point of view, it is important to have a more specific description of the components that build up a system. Many of the laboratory tools used to separate, purify, concentrate, and characterize biomolecules are based on their physical properties such as size, shape, weight, charge and binding properties. These properties are the basis for centrifugation, electrophoresis, dialysis and chromatography (the most powerful methods for fractionating proteins).[2][3]
In addition to identifying the components it is also of key interest to know how they are arranged at the atomic level and what their structure looks like. We are able to deduce a great deal about the electronic and nuclear structure of molecules because of the interaction between electromagnetic radiation and matter. An understanding of basic quantum mechanics (the theoretical explanation for these interactions which form the basis of spectroscopy) allows us to explore the chemical nature of biological systems and their structure: the electronic structure can be studied by

ultraviolet and visible spectroscopy. The rotational and vibrational energy levels (and more importantly, transitions among them) can be observed through infrared spectroscopy. The magnetic spin properties of electrons and nuclei also contain structural information that can be disentangled with nuclear magnetic spectroscopy). Finally, because chemical structures can interact with electromagnetic radiation, twisting and scattering it, we gain structural knowledge from the techniques of polarimetry, circular dichroism, and x-ray crystallography.[2][3]

Biomolecules cannot be properly described by a static picture at room temperature. In general, it is useful to understand how and by which mechanisms the biomolecule evolves over time, so the structure-function studies are the central goal of our investigations. By exploring the details of molecular interactions, we have an understanding of how the observables that we first discovered and characterized, occur.[2]

It is very potent to combine the physical models expressed in mathematical forms along with the experimental structural data collected from thermodynamics and spectroscopy to obtain bona fide structural models with the aid of a computer. A powerful tool available to the structural biologist is the ability to computationally explore potential energy surfaces as biomolecules interact and find their preferred structure in an abstract state space (the space of possibilities). Basic molecular modelling methods (quantum mechanics, molecular mechanics and molecular simulations) accompanied by homology modelling and molecular docking techniques have revolutionized approach to the biomolecule research. With the proper visualization tools and widespread access to many technical servers and databases, computational chemistry has great theoretical and practical appeal.[2][5]

1.2 Molecular interactions

1.2.1 Types of interactions

Molecular interactions – for the reason of gaining shape and forming complexes – are central to structural biology. In general, most of the methods are designed to characterize them at different levels. Interactions can be classified into covalent and non-covalent types.

A covalent interaction is a chemical bond that shares electron pairs between atoms. It allows each atom to obtain the equivalent of a full outer shell, corresponding to a stable electronic configuration.[6] This interaction is characterized by high bond energy, short distances and high directionality, providing irreversible bonding and well defined structure of a molecule.[7] A typical carbon-carbon (C-C) bond has energy of 86 kcal/mol and a length of 1.54 Å. When more than one electron pair is shared between two atoms there is a multiple covalent bond, for example carbon-oxygen (C=O) double bond – it is even stronger than C-C single bond, with energy near 175 kcal/mol and a length of 1.21 Å. Because this energy is relatively high, considerable energy must be expended to break them – chemical reactions entail the breaking and forming of covalent bonds. Covalent bonds hold atoms together within the amino acids and amino acids within proteins (the primary structure). Disulphide bond (60 kcal/mol, 2.05 Å) – usually derived by the coupling of two thiol groups of cysteine residues and playing an important role in the folding and stability of some proteins – is also included to the covalent interactions.[4]

Non-covalent interactions – the forces that drive biology – do not involve the sharing of electrons, but more dispersed variations of electromagnetic interactions between or within molecules. No chemical bonds are formed hence the association is fully reversible; individually, they are weak, but collectively they have a very significant influence on a structure. Furthermore, these interactions are greatly affected in different ways by the presence of water. Non-covalent forces play essential roles in the faithful replication of DNA, the folding of proteins (the secondary and the tertiary structure), the specific recognition of substrates by enzymes, and the detection of molecular signals. They can be generally classified into three categories: electrostatic (including hydrogen bonds and salt bridges), van der Waals and hydrophobic effect.[4][8]

An electrostatic interaction depends on the electric charges on atoms and its energy is given by Coulomb's law – electrostatic potential between partial or full point charges is inversely proportional to the distance separating the charges (Equation 1-1):

$$U_{ij} = k\frac{q_i q_j}{\varepsilon r_{ij}}$$

Equation 1-1. Coulomb's law describing electrostatic interactions. U_{ij} – electrostatic potential, q_i, q_j – electric charges, r_{ij} – distance separating the charges (atoms), ε – dielectric constant, k – conversion factor to the desired energy units.

As an example, the electrostatic interaction between two charged atoms bearing single opposite charges separated by 3 Å in water (which has a dielectric constant of 80) has energy of 1.4 kcal/mol. It is rather long range and it can be an attractive or a repulsive force.[4][7]

Hydrogen bonds are fundamentally electrostatic and energetically weak interactions with certain free energy dependence on directionality. Nonetheless, they are crucial for the function of biological macromolecules and responsible for many of the properties that make water such a special solvent. It forms when a hydrogen atom is positioned between two electronegative atoms, mainly oxygen and nitrogen; the relatively electronegative atom to which the hydrogen atom is covalently bonded pulls electron density away from the hydrogen atom so that it develops a partial positive polarization. Thus, it can interact with an atom having a partial negative polarization (electron density) through an electrostatic interaction. Hydrogen bonds are much weaker than covalent bonds – they have energies of 1-3 kcal/mol compared with approximately 100 kcal/mol for a carbon-hydrogen covalent bond. Hydrogen bonds are also somewhat longer than covalent bonds – their bond distances (measured from the hydrogen atom) range from 1.5 to 2.6 Å, hence, distances ranging from 2.4 to 3.5 Å separate the two non-hydrogen atoms. The strongest hydrogen bonds have a tendency to be approximately straight, such that the hydrogen-bond donor, the hydrogen atom, and the hydrogen-bond acceptor lie along a straight line.[4][7]

Salt bridge is colloquial terms for a bond between oppositely charged residues that are sufficiently close to each other (less than 4 Å) to experience electrostatic attraction. Salt bridges, which can be considered a special form of hydrogen bonds, are composed of negative charges from Asp, Glu, Tyr, Cys and the C-terminal carboxylate group, and of positive charges from His, Lys, Arg and the N-terminal amino group. Since the side chain protonation of these residues is pH dependent, the free energy contributions of salt bridges to protein stability are also pH dependent. The salt bridges of a protein can be grouped according to the number and location of interacting charges. Simple salt bridges are formed between two oppositely charged residues. Salt bridge networks (complex salt

bridges) are composed of three or more interacting charges. Some salt bridges are exposed on the protein surface in an aqueous environment, others are buried in the protein interior, which forms a low dielectric environment. They contribute to protein structure and to the specificity of interaction of proteins with other biomolecules.[9]

Van der Waals interactions exist between all atoms and molecules – its basis is that the electronic charge around an atom changes with time. This transient asymmetry acts through electrostatic interactions to induce a complementary asymmetry in the electron distribution around its neighbouring atoms. The resulting attraction between two atoms increases as they come closer to each other, until they are separated by the van der Waals contact distance (Figure 1-1). At a shorter distance, very strong repulsive forces become dominant because the outer electron clouds overlap. Van der Waals interactions are short range, weak forces – energies associated with them are quite small; typically 0.5-1 kcal/mol per atom pair. When the surfaces of two large molecules come together, however, a large number of atoms are in van der Waals contact, and the net effect, summed over many atom pairs, can be substantial.[4]

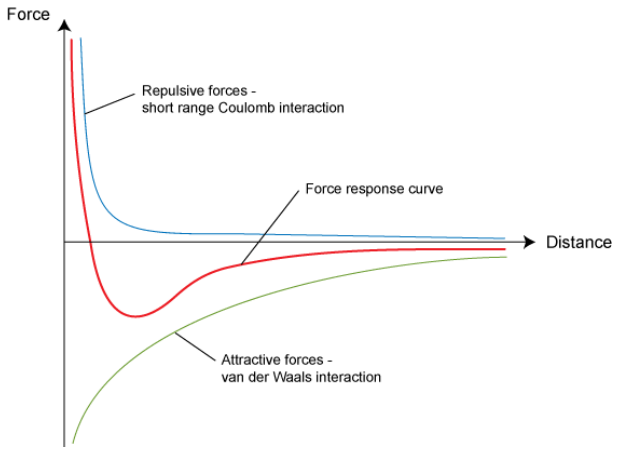


Figure 1-1. Force against distance: the combination of Coulomb and van der Waals interactions results in a force-distance curve approximating the interaction between neutral atoms or molecules, so-called Lenard-Jones potential[10].

Last but not least – hydrophobic effect – the tendency of nonpolar substances to aggregate in aqueous solutions; hydrophilic (polar) regions interact favourably with the solvent and tend to dissolve, but hydrophobic (nonpolar) regions tend to avoid contact with water. Along with the hydrogen bonding, it is the major stabilizing force of conformational stability of folded proteins. It results in the burial of some residues in the core of the protein: 81 % of the nonpolar side chains (Ala, Val, Ile, Leu, Met, Phe, Trp, Cys), 70 % of the peptide groups, 63 % of the polar side chains (Asn, Gln, Ser, Thr, Tyr) and 54 % of the charged side chain (Arg, Lys, His, Asp, Glu) are hidden in the interior of the molecule out of contact with water. Hydrophobic effect among lipids, and between lipids and proteins, is also the most important determinant of structure in biological membranes. Unlike the previously mentioned interactions, the strength of hydrophobic interactions (<10 kcal/mol) is not due to any intrinsic attraction between nonpolar moieties. Rather, it results from the system's achieving greatest thermodynamic stability by minimizing the number of ordered water molecules required to surround hydrophobic portions of the solute molecules.[3][7][11]

1.2.2 Intermolecular interactions

Like any other spontaneous process, a non-covalent binding event takes place only when it is associated with a negative binding free energy (ΔG), which is the well-known sum of an enthalpic term (ΔH) and an entropic term (-TΔS).[12] The free energy of the complex by definition must be lower than the sum of the energies for the solvent separated molecules – in nature, the entropy of a conformation has to be overcome by interactions. The dissociation constant (K_I) also known as the binding constant (K_A) between the components (A, B) of complex (AB) is the ratio of the concentration of the complex divided by the product of the concentrations of the isolated components at equilibrium:

$$A + B \leftrightarrow AB, logK_I = \log\left(\frac{[AB]}{[A][B]}\right) = pK_I$$

Equation 1-2. Dissociation constant characterizing the tendency of a complex to separate reversibly into smaller components.

The types of complexes that form as a result of molecular binding include: protein-protein[13], protein-DNA [14] and protein-ligand (e.g. signalling molecule, drug etc.)[12].

1.2.2.1 Protein-protein interaction characterization

Until the late 1990's, protein analyses mainly focused on single proteins. But because the majority of proteins interact with other proteins for proper function, they should also be studied in the context of their interacting partners to fully understand their intracellular function.[15]
From the partner point of view, two different types can be distinguished, homocomplexes and heterocomplexes. Homocomplexes are usually permanent and optimized. They are responsible for most of the structural cell architecture including tubulin, and the cytoskeleton. Heterocomplexes, in turn, can also have such properties, or they can be no obligatory, being made and broken according to the environment or external factors and involve proteins that must also exist independently.[15]
From the stability point of view, protein interactions are fundamentally characterized as stable or transient, and interactions can be either strong or weak. Stable interactions are those associated with proteins that can be isolated as multi-subunit complexes, and the subunits of these complexes can be identical or different. Hemoglobin and core RNA polymerase are examples of multi-subunit interactions that form stable complexes. Transient interactions correspond to short events and they are expected to control the majority of cellular processes. As the name implies, transient interactions are temporary in nature and typically require a set of conditions that promote the interaction, such as phosphorylation, conformational changes or localization to discrete areas of the cell. Transient interactions can be strong or weak, and fast or slow. While in contact with their binding partners, transiently interacting proteins are involved in a wide range of cellular processes, including protein modification, transport, folding, signalling, and cell cycle regulation.[15]
Proteins bind to each other through a combination of hydrophobic bonding, van der Waals forces, and salt bridges at specific binding domains on each protein. These domains can be small binding

clefts or large surfaces and can be just a few peptides long or span hundreds of amino acids, and the strength of the binding is influenced by the size of the binding domain.[15]
From the experimental point of view, it is difficult to determine complexes or find proper docking algorithms that accurately predict the interaction, despite of the fact that they are essential if we are to understand the kinetics and thermodynamics of complex formation.[13] Usually a combination of techniques is necessary to validate, characterize and confirm protein interactions. A qualitative discovery or verification is already meaningful but a thorough structural characterization is required to understand where, how and under what conditions these proteins interact *in vivo* and the functional implications for these interactions.[15]

1.2.2.2 Protein-ligand interaction characterization

For any two non-identical molecules that associate, it is possible refer to one as the "ligand" and to one as the "receptor". However, most often the receptor will be a protein and the ligand will normally be a smaller molecule. The reason for adopting this terminology lies in the fact that in most practical binding studies, the "receptor" concentration is usually held constant (or nearly so) and the ligand concentration is varied. Moreover, in computational procedures the “receptor” is hold rigid while the ligand is treated as a flexible structure.[16] The main differences between protein-protein and protein-ligand interactions are specificity (selectivity) and cooperativity of the second one. The evolution of protein function includes the development of highly specific sites for the binding of ligands with affinities tailored to meet the needs of a cell. In the same context, cooperativity in ligand binding is usually linked to conformational changes in the protein. Protein-ligand complexes can be properly characterized using well-defined mathematical expressions based on the stoichiometry of the binding equilibrium. Actually, the equilibrium constants provide a thermodynamic measure of the strength of the interaction while the atomic resolution structures of ligand complexes provide the chemical basis for understanding protein–ligand interactions and these structures are often used as the basis for the design of small-molecule drugs for the treatment of disease.[17]

In the presented work, protein-protein interactions have been studied on the example of connexin proteins and protein-ligand interactions have been studied on the example of two heme pathway enzymes; both cases introduced with details in the following sections, 1.3 and 1.4, respectively.

1.3 Connexin as the key elements in the direct cell-cell communication process

1.3.1 Cell-cell communication

Cells do not live in isolation; each cell has been programmed during development to communicate with other cells. The molecule and information transfer requires a complementary set of channel or receptor proteins that enable it to receive and respond to the signals in a characteristic way.[18]
This process not only promotes the proper functioning of individual cells, but it also allows communication and coordination among groups of cells — including the cells that make up

organized communities, called tissues. Thanks to cell signalling and junction, tissues have the ability to carry out tasks that no single cell could accomplish on its own. To preserve proper tissue architecture, adhesive molecules help maintaining contact between nearby cells and structures, and tiny tunnel-like junctions allow the passage of ions and small molecules between neighbouring cells. Meanwhile, signalling molecules relay information among the cells in a tissue, as well as between these cells and the extracellular matrix. These pathways are critical to maintain the state of equilibrium known as homeostasis within an organism.[19]

With the exception of a few terminally differentiated cells such as skeletal muscle cells and blood cells, most cells in animal tissues are in communication with their neighbours via gap junctions (GJs). GJs are the only known cellular structures that allow direct transfer of small molecules (ions (K^+ and Ca^{2+}), second messengers (cAMP, cGMP, and inositol 1,4,5-triphosphate (IP_3)), and small metabolites (glucose)) by forming hydrophilic protein channels that bridge the membranes of adjacent cells. Assembly and functioning of these channels are complex and highly regulated processes, to date not well understood.[18][20]

1.3.2 Gap junction organization and connexin characterization

In mammals, GJ channels are composed of two head-on docked hemichannels (connexons) provided by two neighbouring cells (Figure 1-2). Each connexon is composed of 6 protein subunits (connexins (Cx)) belonging to a multigene family with more than twenty classes identified.[21] There are two conventions for connexin nomenclature – one depends on molecular mass of a subunit (Cx26 represents the connexin protein of 26 kDa; also hemichannels further made), the other uses Greek symbols based on evolutionary considerations (GJB2 stands for gap junction beta 2 referring to Cx26).[20]

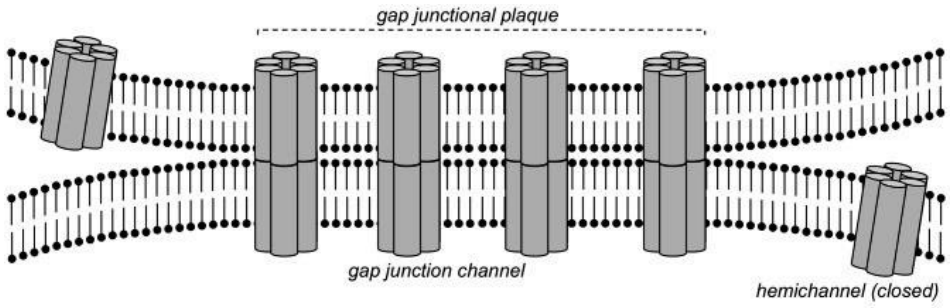


Figure 1-2. Gap junction channels. Schematic representation of a gap junctional plaque, with each of the neighbouring cells contributing one hemichannel to form the complete gap junction channel.[22]

Cxs are synthesized and assembled to connexons in the endoplasmic reticulum (ER) membrane, later on they are transported to the cell surface in vesicular carriers that travel through Golgi, along microtubules, following the general intracellular transport route referred to as the secretory pathway (Figure 1-3). Hemichannels are distributed over the entire surface and they can move laterally in the plane of the membrane reaching the outer margins of GJ plaques. Typically a single plaque may contain between less than a dozen to thousand individual channels and it can extend from less than a hundred nanometres to several micrometres in diameter. Different Cxs show disparate preferences for certain lipid environments and it has been hypothesized that these

preferences can be involved in channel aggregation, mixing, and segregation. In some cases, Cx isoforms co-distribute homogenously while in other arrangements they are segregated into well-separated domains. GJ channels, once formed, cannot be separated again into hemichannels under physiological conditions. The half-life of Cx, in general ranges from about 1 to 5 h both *in vivo*, and *in vitro*. Its degradation appears to involve both the lysosomal and proteasomal degradation pathways.[21]

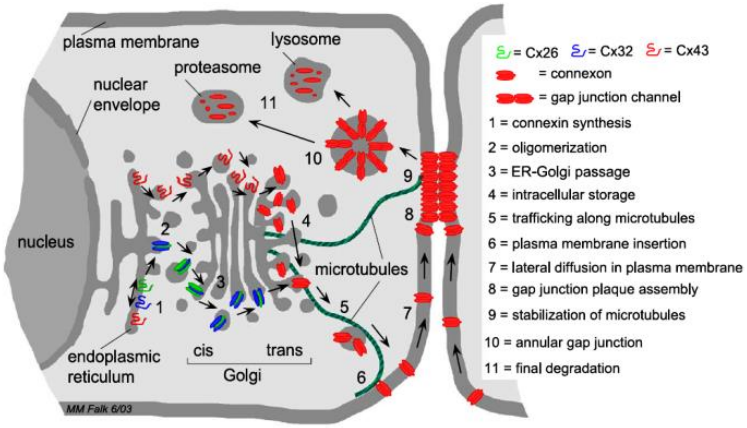


Figure 1-3. Gap junction channel path of life. Schematic representation of the steps that lead to synthesis, assembly, and degradation of GJ channels.[21]

Nearly all cells in the body express at least one type of Cx isoform at some point during development or within adult life.[20] Different types of Cxs lead to different types of channels with ultimately different biophysical properties of conductance, gating and permeability. Hemichannels can be formed either from a single type of Cx (homomeric) or from more than one type (hetromeric). In turn, homotypic channels can be formed from either the same homomeric or the same heteromeric connexons, whereas heterotypic channels are those that connect different homomeric or heteromeric hemichannels (Figure 1-4).[22]

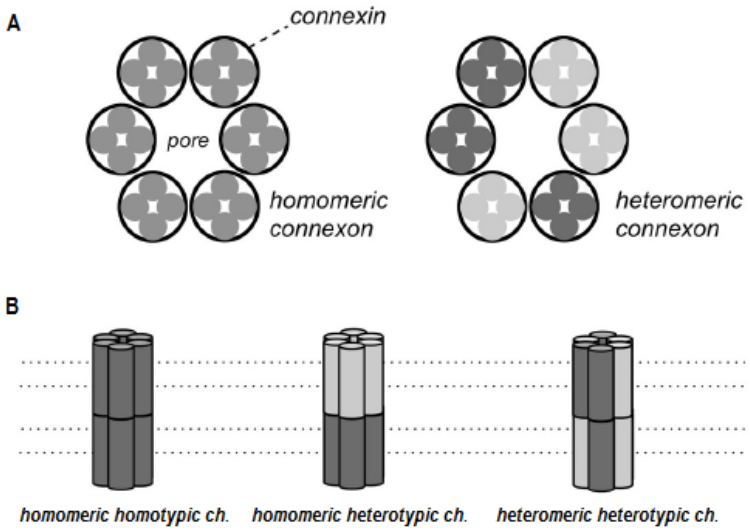


Figure 1-4. Connexin organization in gap junction channels. (A) Connexon types: homomeric (left), heteromeric (right). (B) Channel types: homomeric homotypic (left), homomeric heterotypic (centre), heteromeric heterotypic (right).[22]

Assuming that each connexin monomer within a junctional channel contributes to some degree to the channel physiology, for two connexins, there are 5 possible heteromeric stoichiometry and 12 possible arrangements within a hemichannel. Adding the 2 homomeric possibilities, a total of 14 structural hemichannel forms are possible.[23] The mere existence of functional heteromeric and heterotypic channels indicates a high degree of molecular recognition between different connexin chains, strongly suggesting a high structural homology. However, the number of theoretical combinations of connexin isoforms far exceeds the actual number of different channels which have been observed/reported (Table 1-1) and later on divided into compatible groups (Table 1-2). Functional heterotypic gap junction channels can be formed among one group (A), or the other (B) but not between them.[23]

Cx26	Cx30	Cx30.3	Cx31	Cx31.1	Cx32	Cx33	Cx36	Cx37	Cx40	Cx43	Cx45	Cx46	Cx50	Cx57	
+	+	–	–	–	+		–	–	–	–	–	+	+	–	Cx26
	+	+			+		–	+	+	+	+			–	Cx30
		+	–	–	–		–	+	+	+	–			+	Cx30.3
			+	–	–			–	–	–	–			–	Cx31
				–	–			–	–	–	–			–	Cx31.1
					+	–	–	–	–	–	–	+	+	–	Cx32
						–		–		–					Cx33
							+	–	–	–	–		–		Cx36
								+	+	+	+			+	Cx37
									+	+	+	–	–	–	Cx40
										+	+	+	–	+	Cx43
											+			–	Cx45
												+	+	–	Cx46
													+	–	Cx50
														+	Cx57

Table 1-1. Heterotypic functional compatibilities among mammalian connexins; +, forms heterotypic channels – , does not form heterotypic channels, blank – not reported. [23]

Group A	Group B
Cx30.3	Cx26
Cx37	Cx32
Cx40	Cx46
Cx43	Cx50
Cx45	
Cx57	

Table 1-2. Heterotypic functional compatibility groups.

The compatibility, stability and regulatory determinants for the connexin selection in the cell and later during the docking process are not known. There are several possible mechanisms, not completely understood: (i) different Cx isoforms may be synthesized in different regions of the ER, and thus, will not get into physical contact to allow their heteromerization; (ii) specific chaperones may bind to particular Cxs and prevent their interaction; (iii) in analogy to ion channels, specific signals (structural motifs) encoded in the connexin polypeptides may regulate Cx interactions, and allow oligomerization of only compatible isoforms.[21]

Albeit its importance, there is only limited structural information for the GJs. Until now the structure of Cx26 at 3.5 Å resolution[24] is the only one available, and may serve as a basis for the computational modelling of other Cxs (Figure 1-5 A,B). Four transmembrane domains generate the hemichannel pore (TM1, TM2, TM3, TM4), two extracellular loops have been associated with cell–cell recognition and docking (E1, E2) and the cytoplasmic moieties (N and C termini plus the cytoplasmic loop; NT, CT, CL) are believed to have regulatory roles (Figure 1-5 C).[20]

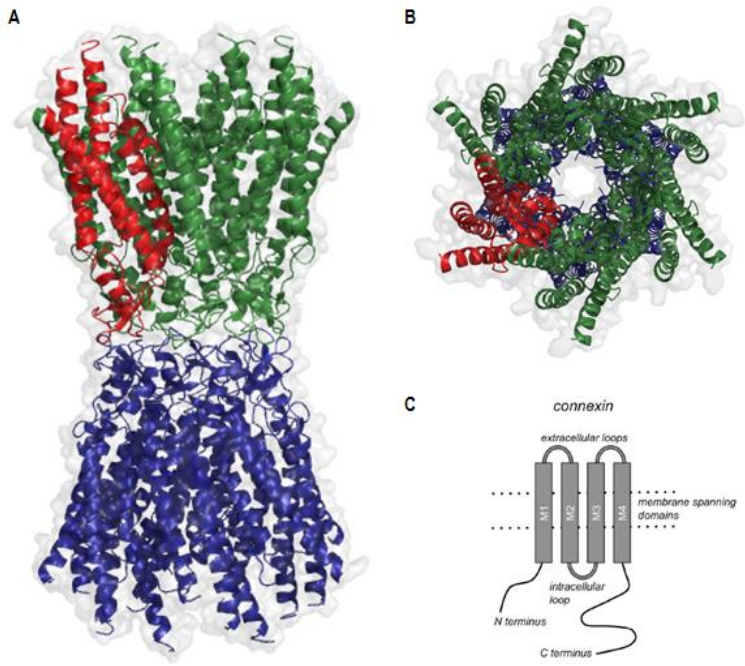


Figure 1-5. Overall structure of a gap junction channel. (A) Side view of the Cx26 GJ channel. (B) Top view of the Cx26 GJ channel. (C) Structure of the pattern of a single connexin subunit with 4 domains traversing the membrane.[24][22]

Genetic variations in the primary sequence of Cxs are directly related with disease, with almost 700 mutations in genes carrying Cxs described, causing a variety of pathologies, such as skin disorders, hearing loss or epilepsy.[25,26] Mutations can impair protein trafficking and/or the connexon assembly at the membrane, they can constitutively affect the active hemichannel architecture or they can alter the range of available channel types through a mechanisms called *trans*-dominant inhibition that, in ultimate instance, alters the type of molecules that are exchanged and the extent of intracellular communication.[20] Confocal imaging and patch-clamp electrophysiology experiments allow quantifying the loss intracellular communication produced by the mutation, but the structural bases and the mechanism of action for such deleterious mutants are still poorly understood.

1.4 Porphobilinogen deaminase and Uroporphyrinogen III Synthase as components of the heme biosynthetic pathway

1.4.1 Tetrapyrroles in heme pathway and porphyrias

Tetrapyrroles like hemes, chlorophylls, and cobalamin are complex macrocyclic molecules which play essential roles in almost all living organisms and take part on several metabolic processes, including electron transfer during respiration, photosynthesis, and enzyme catalysis. In this context, the iron-containing porphyrin heme fulfils widely diverse biological functions as prosthetic group for many enzymes, transporters, and receptors in most living organisms. Heme-containing cytochromes are integral components of various respiratory and photosynthetic electron transport chains. Heme serves as a prosthetic group in hemoglobin, myoglobin, catalases, peroxidases, cytochromes P450, and in sensor proteins for diatomic gases such as O2 or NO. In all eukaryotes and most prokaryotes, except the archaea and some eubacteria, heme is synthesized via a well-

established and conserved biosynthetic pathway, accomplished by the sequential action of eight enzymes, in human body – mainly in the liver and in the erythroid cells. Malfunction of this pathway (decreased activity of enzymes due to genetic defects) leads to severe metabolic disorders, rare disease family termed porphyrias. The type of porphyria depends on the specific enzyme affected by a mutation (Figure 1-6).[27–30]

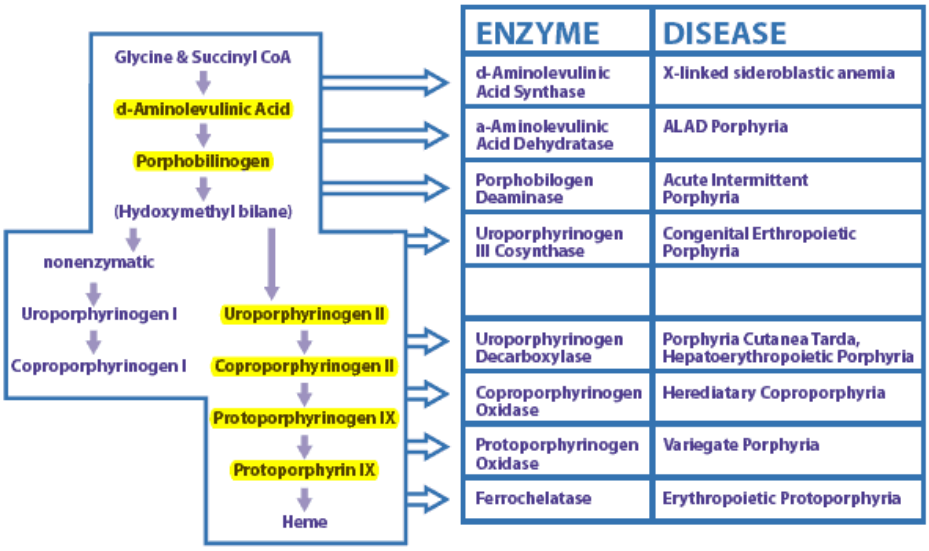


Figure 1-6. Diseases due to deficiencies of specific enzymes of the heme biosynthetic pathway.[31]

The biosynthesis of heme sets off with the formation of D-aminolevulinic acid (ALA) as the common precursor for all naturally occurring tetrapyrroles. ALA represents the sole source of carbon and nitrogen atoms necessary for heme formation. Two ALA molecules are then condensed to form the pyrrole porphobilinogen (PBG). Oligomerization of four PBG molecules yield the linear tetrapyrrole intermediate pre-uroporphyrinogen (1-hydroxymethylbilane (HMB)) followed by ring closure to lead the cyclic tetrapyrrole intermediate uroporphyrinogen III (URO III). Subsequently, the side chains of the macrocycle are modified resulting in the intermediates coproporphyrinogen III and protoporphyrinogen IX. This is followed by the aromatization of the ring system to produce protoporphyrin IX. Finally, iron is inserted to yield heme. In summary, heme biosynthesis can be divided into three parts – (i) formation of the precursor molecule ALA, (ii) formation of the first cyclic tetrapyrrole URO III, and (iii) conversion of URO III into heme – and it requires a highly equilibrated cooperation between a number of enzymes to achieve efficient transfer of metabolites. Several intermediates on the biosynthetic road to heme are highly unstable or reactive molecules. Furthermore, all porphyrinogens are easily oxidized to the corresponding porphyrins and these in turn are extremely sensitive to light. Such compound instability is a source of problems for the porphyria patients, for whom one of the consequences of the pathological accumulation is irreversible tissue damage. [27,32]

URO III represents the first cyclic intermediate of tetrapyrrole biosynthesis and, at the same time, the first major branch point for the diverging routes leading to formation of different classes of tetrapyrroles. The crucial part of the conversion of PBG into URO III depends upon cooperative action of two enzymes – Porphobilinogen Deaminase (PBGD) and Uroporphyrinogen III Synthase (UROIIIS).[27,32,33]

1.4.2 Porphobilinogen deaminase (PBGD)

Different methods – structural and biochemical – have been used to characterize PBGD. [34–37] The most interesting feature for this enzyme is the multistep reaction it catalyses – formation of a linear tetrapyrrole product, pre-uroporphyrinogen, from four units of porphobilinogen (PBG) (Figure 1-7).

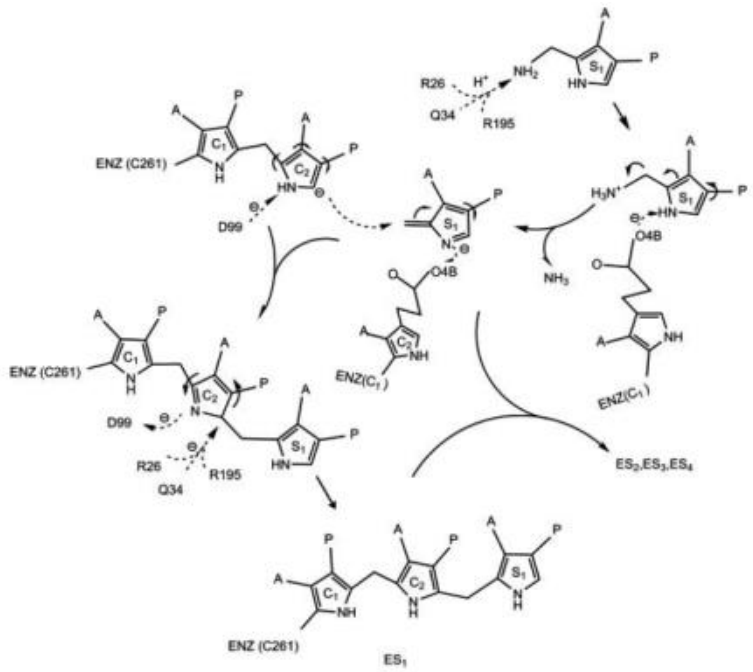


Figure 1-7. Schema of the PBGD substrate catalysis mechanism. PBG (S1) is deaminated to methylene pyrrolinene, and the free α-position of the C2 ring of the DPM cofactor, on activation by D99, forms a carbon-carbon bond with the methylene pyrrolinene. After stabilization of the intermediate ES1, the reaction proceeds through ES2, ES3, and ES4 stages, with the addition of a pyrrole ring at each stage in a similar fashion. ES4 is then hydrolysed, releasing the product, regenerating the enzyme with the DPM cofactor still bound at the active site.[34]

PBGD uses a single active site yet it repeats the same action four times, while the product – attached to the dipyrromethane cofactor (DPM) – is being extended. This covalently attached unique cofactor consisting of two linked PBG molecules serves as a primer for the oligomerization of another PBG molecules resulting in the formation of a protein bound linear hexapyrrole. From this precursor the linear tetrapyrrole pre-uroporphyrinogen is then cleaved off leaving the free enzyme with its cofactor behind.[38] Till now, it is very unclear how exactly the enzyme deals with the growing oligopyrrole chain during catalysis and the cleavage of the product.

PBGDs, existing as monomers with a mass ranging from 34 to 45 kDa, are highly conserved across different species. Human PBGDs share over 45% primary sequence identity with their counterparts from *E. coli* and others[35]. At least two isoforms have been reported: the 44-kDa ubiquitous (housekeeping) enzyme and the 42-kDa erythrocyte-specific enzyme[39]; they cannot be confused with reaction intermediates). The ubiquitous PBGD (represented by the deposited human PBGD structures) has 17 extra residues in the N-terminal region when compared to erythrocytic PBGD. It is not clear what, if at all, is the role of the N-terminal segment. To date, ten crystal structures of the protein are available in the Protein Data Bank (2 of *H. sapiens*, 5 of *E. coli*, 2 of *B. megaterium* and

1 of *A. thaliana*). However, the structures of subsequent catalytic stages of PBGD that are essential to understand the mechanism of the reaction are still missing.

The structure of PBGD consists of three equally sized α/β domains connected by three hinge regions and a segmental domain motion regions those regions has been predicted as a mechanism to allocate the sequential entry of PGB molecules. The active site is located at the interface of domains 1 and 2 (Figure 1-8 A). The DPM cofactor is covalently bound to Cys261 (human protein numbering) via a thioether linkage and is sitting in a positively charged cleft region, which is large enough to accommodate more molecules of methylene pyrrolinene during catalysis (Figure 1-8 B). On the other hand, probable steric considerations prevent the addition of more rings beyond the ES4 tetrapyrrole stage.[34]

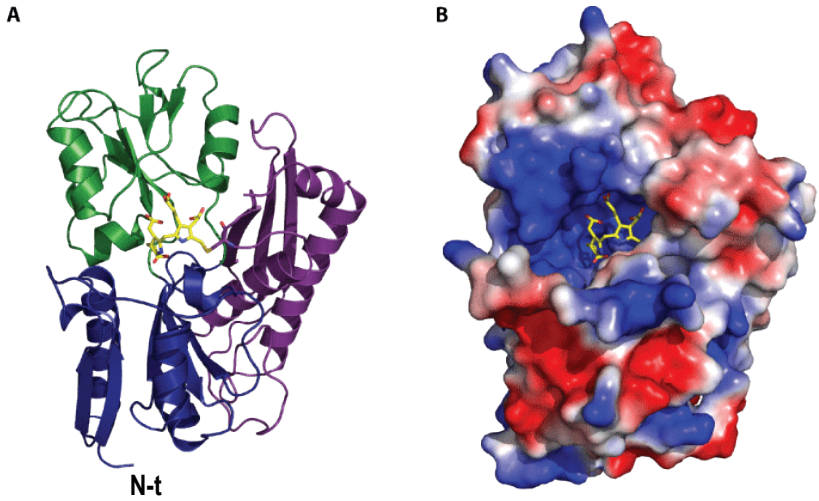


Figure 1-8. Overall structure of human PBGD. (A) Cartoon representation bound with DPM cofactor; dark blue – domain 1, green – domain 2, purple – domain 3, N terminal of the protein marked as N-t. (B) Surface electrostatic potential representation; blue – positive potential, red – negative potential. Cofactor shown as sticks. Resolution: 2.2 Å. [34]

A number of residues are forming hydrogen bonds with the cofactor and substrates, more precisely – with the carboxyl oxygens of their acetate and propionate side chains. The interactions, mainly with positively charged residues (Arg26, Gln34, Lys98, Arg149, Arg150, Arg173, Arg95, Gln217), are predominantly ionic and most probably involved in the correct positioning of incoming molecules. Interestingly, all these residues are totally conserved in PBGDs from different species, suggesting a highly conserved mechanism of substrate binding and catalysis. The hypothesized key residue for catalysis is Asp99 that may play an important role by assisting in the substrate deamination.[34,40,41]

1.4.3 Uroporphyrinogen III Synthase (UROIIIS)

UROIIIS follows PBGD in the pathway, catalysing the cyclization of the linear pre-uroporphyrinogen (HMB) to the macrocyclic URO III under inversion of a ring D (Figure 1-9). A catalytic mechanism has been proposed in which rearrangement of the A ring of HMB results in the loss of the C20 hydroxyl group. Created carbo-cation at C20 performs an electrophilic attack on C16 of D ring forming an spiranic intermediate which next cyclizes to URO III.[42] [43]

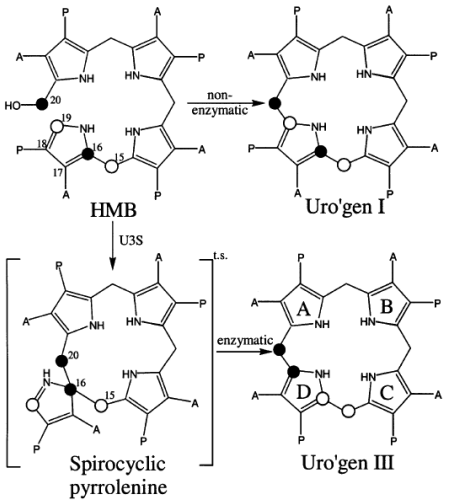


Figure 1-9. Schema of the mechanism of catalysis of UROIIIS. Conversion of HMB to URO I (non-enzymatic, here marked Uro'gen I) and URO III (enzymatic, marked as Uro'gen III). A – acetate, P – propionate, spirocyclic pyrrolenine – proposed transition state intermediate.[43]

The domains, show similar folds that comprise a parallel β-sheet surrounded by α-helices. However, they are not the same fold types; domain 1 belongs to the flavodoxin-like fold family and domain 2 adopts a DNA glycosylase-like fold. The linker connecting the two domains adopts an antiparallel β-ladder conformation in the human UROIIIS structure.[43]

The distribution of conserved residues suggests that the enzyme active site is located in the large open cleft between domains. Point mutation experiments show that the enzymatic mechanism does not include acid/base catalysis. Unlike PBGD, UROIIIS does not make extensive use of charge–charge interactions with the carboxylate side chains, as only one of the conserved residues is positively charged. Instead, variations in the inter-domain orientations observed in the *T. thermophilus* UROIIIS crystallized at several different conformations indicate the presence of flexibility (opening and closing) that may be important in the catalytic cycle; the linker seems to be highly mobile in solution in the absence of the substrate. Concluding, all evidences point to the fact that a large number of residues are involved in the catalytic mechanism.[43,44] Most of the known mutations disrupt the overall structural integrity of the enzyme leading to the reduction of its stability and thus activity.[45]

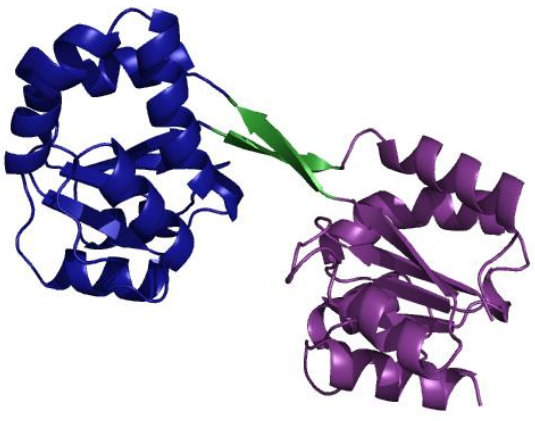


Figure 1-10. Overall structure of human UROIIIS. Cartoon representation; purple – domain 1, dark blue – domain 2, green – linker, 1.85 Å resolution.[43]

The domains, show similar folds that comprise a parallel β-sheet surrounded by α-helices. However, they are not the same fold types; domain 1 belongs to the flavodoxin-like fold family and domain 2 adopts a DNA glycosylase-like fold. The linker connecting the two domains adopts an antiparallel β-ladder conformation in the human UROIIIS structure.[43]
The distribution of conserved residues suggests that the enzyme active site is located in the large open cleft between domains. Point mutation experiments show that the enzymatic mechanism does not include acid/base catalysis. Unlike PBGD, UROIIIS does not make extensive use of charge–charge interactions with the carboxylate side chains, as only one of the conserved residues is positively charged. Instead, variations in the inter-domain orientations observed in the *T. thermophilus* UROIIIS crystallized at several different conformations indicate the presence of flexibility (opening and closing) that may be important in the catalytic cycle; the linker seems to be highly mobile in solution in the absence of the substrate. Concluding, all evidences point to the fact that a large number of residues are involved in the catalytic mechanism.[43,44] Most of the known mutations disrupt the overall structural integrity of the enzyme leading to the reduction of its stability and thus activity.[45]

1.4.4 Porphyrias

An enzyme acts by reducing the energy barrier in the pathway from a chemical (substrate) to form a product. In any biosynthetic pathway the product of one enzyme is the substrate for the next one. When an enzyme is deficient, the substrate for that enzyme will accumulate if it does not catabolize well. Deficiencies in the activities of the enzymes of the heme biosynthetic pathway – as it was mentioned in the 1.4.1 section – cause uncommon and complex metabolic conditions known as porphyrias. Each porphyria is due to a deficiency (but with some residual activity) of one of the enzymes of the hem group biosynthetic pathway (Figure 1-6). Depending on the type of porphyria, the excess intermediates are produced and accumulate first in the liver or the bone marrow and then appear in blood and excreted from the body in urine and feces. Porphyrias manifest with diversity of symptoms, from neurological complications to skin problems. PBGD and UROIIIS mutations are responsible for, respectively, the acute intermittent porphyria (AIP) and the congenital erythropoietic porphyria (CEP).[30,31]
Acute intermittent porphyria (AIP) is the most common of the porphyrias affecting the liver (prevalence of about 1 in 20 000 people), inherited as an autosomal dominant disorder. Most affected individuals are heterozygotes and are usually asymptomatic, because the single copy of the WT-coding gene provides a sufficient level of enzyme function. AIP manifests with diverse group of symptoms - abdominal pain (85-95 %), vomiting (50 %), constipation (50 %), peripheral neuropathy (42-68 %), seizures (10-16 %), delirium, coma and depression - making its early diagnosis difficult. Delayed diagnosis and treatment can be fatal or can cause long term or permanent neurological damage. The attack is often precipitated by environmental factors, reduced calorie intake, medications (barbiturates, calcium channel blockers, antibiotics, antifungals, and hormones), by a large alcohol intake, nicotine abuse, infection, surgery and by psychiatric illness. Characteristic features are elevated levels of ALA and PBG in the patients' urine and a quantitative measurement of such metabolites constitutes the most reliable confirmatory tests. Treatment of

AIP during acute attack intends to decrease the heme synthesis and production of porphyrin precursors with removal of precipitating factors, treatment of underlying infection and a carbohydrate diet. In severe cases heme therapy is indicated. Heme acts by depressing the ALA synthetase enzyme and it is available in the form of hematin, heme albumin or heme arginate. Between attacks patients are usually healthy.[46–48] In total, there are more than 100 known mutations causing AIP in humans either due to less stable or truncated forms of the protein or defects in cofactor and substrate binding.[27]

On the other hand, mutations in human UROIIIS gene (more than 45 identified [49]) cause congenital erythropoietic porphyria (CEP), autosomal recessive disorder. Reduced UROIIIS activity leads to the nonenzymatic conversion of highly unstable HMB to nonphysiologic, functionally irrelevant yet toxic symmetric isomer – URO I (Figure 1-9). URO I, followed by coproporphyrin I and other symmetric derivatives, accumulate in late erythroid precursors, reticulocytes, and red cells, resulting in ineffective erythropoiesis, hemolysis, and splenomegaly. They disseminate into tissues such as skin to cause devastating photosensitivity; photo excited porphyrins are quite reactive. Also because of large amounts of porphyrins excretion and deposition the urine of patients is red and their teeth exhibit a strong red fluorescence under ultraviolet light.[4,50]

CEP – also known as Günther's disease – is extremely rare (prevalence estimates of 1 in 1 000 000 or less), still it is the most frequent of the recessive porphyrias. In contrast to AIP, CEP does not cause acute attacks but is a permanent disorder. Its severity correlates with the degree that enzymatic activity is reduced and its management is largely comprised of sun avoidance, dietary changes, continuous administration of heme derivatives or transfusing blood to counteract heme group deficiency, and supportive measures to prevent general aggravation. However, effective treatment have not been established yet.[47,48]

Therapies such as hypertransfusion[51,52] and oral charcoal[53–55] have had limited success. Iron chelation[56] or its induced deficiency[50] has been used to ameliorate the accompanying iron overload that often develops in CEP patients. Other efforts have focused on bone marrow transplantation[57,58] and gene therapy - under development and has not yet been tested on humans[59]. From the examples, the only curative treatment, performed successfully in a small number of cases, is a bone marrow transplant where photosensitivity or anemia problems were resolved. However, not the scars from previous skin lesions. Difficulties are associated with the two facts: (i) for a successful transplant the bone marrow must be of very high similarity between a donor and a recipient, (ii) powerful treatments inhibiting the recipient's immune system are initially required to prevent rejection. Same, a bone marrow transplant is reserved for those severely affected individuals having an identical bone marrow donor.[60] Hence, there is an urgent need of alternative treatment options that increase the quality of life of people affected by CEP.

1.4.5 Inhibition and drug designing as a perspective for porphyria therapies

Enzymes, like PBGD and UROIIIS, are molecular devices that possibilitate chemical transformations sustaining life of a cell. By utilizing the full repertoire of intermolecular forces, they bring distinct compounds together in an optimal orientation, creating and disrupting chemical bonds. The unique features of enzymes are their catalytic power and specificity with respect to the

place and the substrate. They are powerful because they increase the rate of a reaction by lowering its activation energy and as a result, products are formed faster and reaction reach the equilibrium state more rapidly .[4]

The kinetic properties of most of enzymes can be characterized by the Michaelis-Menten model (Equation 1-3). In this model, an enzyme (E) combines with a substrate (S) to form an enzyme-substrate (ES) complex, which can proceed to form a product (P) or to dissociate into E and S:

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \overset{k_2}{\underset{X}{\rightarrow}} E + P$$

Equation 1-3. Michaelis-Menten model for description of the kinetic properties of enzymes; E – enzyme, S – substrate, P – product, k – reaction kinetic constants.

The product formation rate is given by the Equation 1-4:

$$V_0 = \frac{d[P]}{dt} = V_{max} \frac{[S]}{[S] + K_M}$$

Equation 1-4. Michaelis-Menten equation. Vo - catalytic rate of product formation, Vmax – maximal reaction rate, KM – Michaelis constant.

The maximal rate, Vmax, is the reaction rate when the enzyme is fully saturated with substrate and is equal to the product of k2 or kcat and the total concentration of enzyme. The kinetic constant kcat, called the turnover number, is the number of substrate molecules converted into product per unit time at a single catalytic site when the enzyme is fully saturated with substrate. Turnover numbers for most enzymes are between 1 and 104 per second. The ratio of kcat/KM provides a penetrating probe into enzyme efficiency. KM, the Michaelis constant, is the substrate concentration at which the reaction rate is half maximal.[4]

Enzymes, although highly specific, are not completely independent. Some of them require cofactors (small molecules like vitamins or metal ions) for their catalytic activity. Some of the enzymes may be inhibited (stopped) by other small molecules and this phenomenon serves as a major control mechanism in biological systems. Many drugs and toxic agents act by inhibiting enzymes. Inhibition by particular chemicals can also be a source of insight into the mechanism of enzyme action: specific inhibitors are often used to identify residues critical for catalysis.[4]

There are many types of enzyme inhibition yet, we am going to describe only those that are most important for the current project: competitive and non-competitive inhibition (Figure 1-11).

In competitive inhibition, an enzyme can bind substrate (forming an ES complex) or inhibitor (EI) but not both simultaneously (ESI). A competitive inhibitor typically resembles the substrate and binds to the active site of the enzyme. The substrate is thereby prevented from binding to the same active site. A competitive inhibitor diminishes the rate of catalysis by reducing the proportion of enzyme molecules bound to a substrate. At any given inhibitor concentration, competitive inhibition can be reverted by increasing the substrate concentration. In non-competitive inhibition, which also is reversible, the inhibitor and the substrate can bind simultaneously to an enzyme molecule yet at different binding sites. A non-competitive inhibitor acts by decreasing the turnover number rather than by diminishing the proportion of enzyme molecules that are bound to

substrate. Non-competitive inhibition, in contrast with competitive inhibition, cannot be overcome by increasing the substrate concentration.[4]

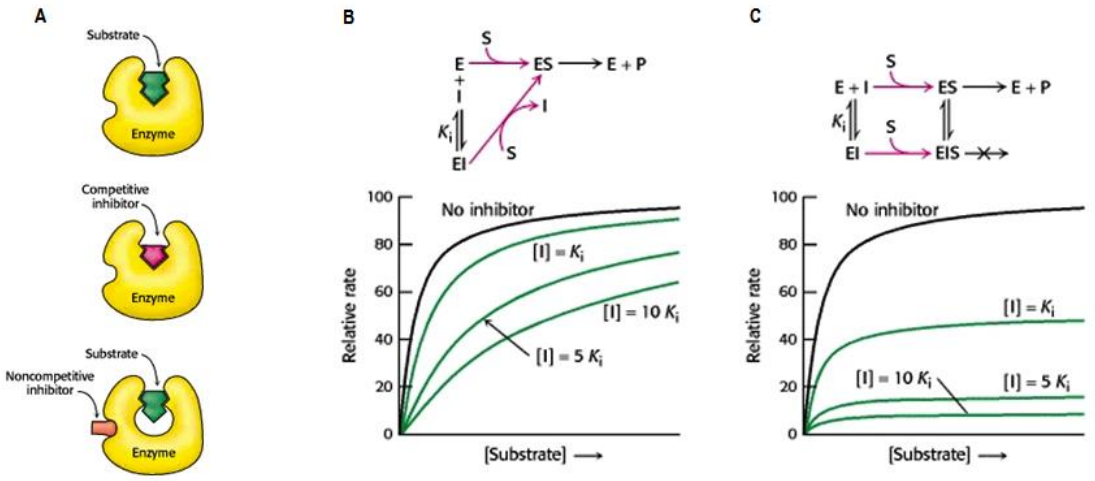


Figure 1-11. (A) Distinction between a competitive and a non-competitive Inhibitor: (top) enzyme-substrate complex; (middle) a competitive inhibitor binds at the active site and thus prevents the substrate from binding; (bottom) a non-competitive inhibitor does not prevent the substrate from binding. (B) Kinetics of a competitive inhibitor: As the concentration of a competitive inhibitor increases, higher concentrations of substrate are required to attain a particular reaction velocity. The reaction pathway suggests how sufficiently high concentrations of substrate can completely relieve competitive inhibition. (C) Kinetics of a non-competitive inhibitor: The reaction pathway shows that the inhibitor binds both to free enzyme and to enzyme complex. Consequently, Vmax cannot be attained, even at high substrate concentrations.[4]

Competitive and non-competitive inhibitions are kinetically distinguishable by the measurements of the rates of catalysis at different concentrations of substrate and inhibitor. In competitive inhibition, the inhibitor competes with the substrate for the active site and the dissociation constant for the inhibitor is given by the Equation 1-5:

$$K_i = \frac{[E][I]}{[EI]}$$

Equation 1-5. Dissociation constant for the enzyme-complex (Ki), I – inhibitor.

Because increasing the amount of substrate can overcome the inhibition, Vmax can be attained in the presence of a competitive inhibitor (Figure 1-11 B). However, the apparent value of KM is altered, increased:

$$K_M^{app} = K_M\left(1 + \frac{[I]}{K_i}\right)$$

Equation 1-6. Apparent Michaelis constant ($\mathbf{K_M^{app}}$); competitive inhibition.

In non-competitive inhibition (Figure 1-11 C), a substrate can still bind to the enzyme-inhibitor complex but the enzyme-inhibitor-substrate complex does not proceed to form product. The value of Vmax is decreased to a new apparent value (Equation 1-7), while the value of KM is unchanged – the inhibitor simply lowers the concentration of functional enzyme and the remaining enzyme behaves like a more dilute solution of enzyme.[4]

$$V_{max}^{app} = \frac{V_{max}}{1 + \frac{[I]}{K_i}}$$

Equation 1-7. Apparent maximal reaction rate; non-competitive inhibition.

As mentioned before, most drugs produce their effect by inhibiting enzymes. The interaction between a ligand and its target (receptor) usually is due to non-covalent forces, but in some cases a covalent interaction may also be involved; the ligand often forms hydrogen bonds with the receptor, also some receptors have hydrophobic "pockets", formed by groups of non-polar amino acids, that recognize the ligand through a hydrophobic group of an appropriate size. Strong ligand binding often shows a high degree of complementarity with the region of the target molecule where it binds. Searching for this complementarity, measuring it and characterizing in various ways is a first step in the discovering any new drug. Two key steps are the identification of hit molecule that has some reproducible activity in a biological assay and lead series that comprises a set of related molecules that share some common structural feature – from among them it is expected to find a drug candidate with the desired response, potency and selectivity (enabling it to reach its target in vivo), lack of toxicity and minimal side-effects. Such a drug candidate will then enter the long and expensive process of development, where further large-scale investigation (including clinical phases) is undertaken. In the fight to reduce costs and workload, computational chemistry has become a perfect ally with various techniques being available depending upon type of molecule to screen, the kind of information available to assist on the selection and the properties to take into consideration. A wide variety of methods can be used either individually or in combination – virtual libraries of molecules allow the fast screening for compounds, different docking methods, based on different scoring functions to facilitate searching and characterizing the protein active sites. From the point of view of a reaction or inhibition, the activity modelling employs tools such as energy minimisation, molecular dynamics, Monte Carlo simulations and conformational analysis. Many different strategies are possible but the goal is always to transfer to the practical trials a set of compounds, carefully selected and with the most promising properties.[5]

The CEP therapy quest, in which our group is involved, often aims at reducing levels of circulating porphyrins as the disease symptoms are highly correlated with the degree of its excess.[61]

Our group propose (and work on) two different paths of investigation: (i) reduction of the amount of pre-uroporphyrinogen by the inhibition of PBGD and therefore cutting UROIIIS off from the substrate supplies, (ii) use of molecular chaperones which bind to UROIIIS and stabilizes its folded (active) conformation, pulling back mutant protein's intracellular degradation. The steps in which I was involved during the course of my doctorate studies are described in the following sections: initial phase of designing and characterizing potential PBGD inhibitors – 4.2.1, inspecting the structural basis for UROIIIS destabilization by a hotspot mutation as the first stage for chaperone-based therapies – 4.2.3.with details in the following chapters, 1.3 and 1.4, respectively.

OBJECTIVES

2

2.1 Protein-protein interactions & connexins

- To characterize the protein-protein interactions between connexin hemichannels by studying Cx26, Cx32 and Cx47 docking process (in comparison with experimental results).

- To characterize the effect of the mutation on the hemichannel docking process and proper functioning.

- To characterize the role of different structural motifs of connexin chains using molecular docking and molecular dynamics simulations.

2.2 Protein-ligand interactions & heme pathway enzymes

- To evaluate *in silico* the potency of PBGD inhibitors.

- To determine the residues of a heme pathway enzyme, PBGD, that are key for catalysis, using molecular docking and mutagenesis.

- To separate and structurally characterize the different PBGD reaction intermediates.

- To characterize the stability of the most common mutation of UROIIIS using computational methods.

MATERIALS & METHODS

3

3.1 Computational methods

In the field of molecular modelling, defined as a collection of theoretical and computational techniques used to mimic the behaviour of a system, we can distinguish three main streams: i) **energy description** (quantum mechanics (QM) – *ab initio* and density functional methods (DFT), semi-empirical methods and molecular mechanics (MM). ii) **trajectory description** (molecular dynamics (MD) and Monte Carlo method) and iii) **approach description /prediction of molecular complexes** (docking simulations).

3.1.1 Sequence and structure groundwork

To be able to run any kind of simulation, a reliable structure coordinate file is needed.

3.1.1.1 Homology modelling

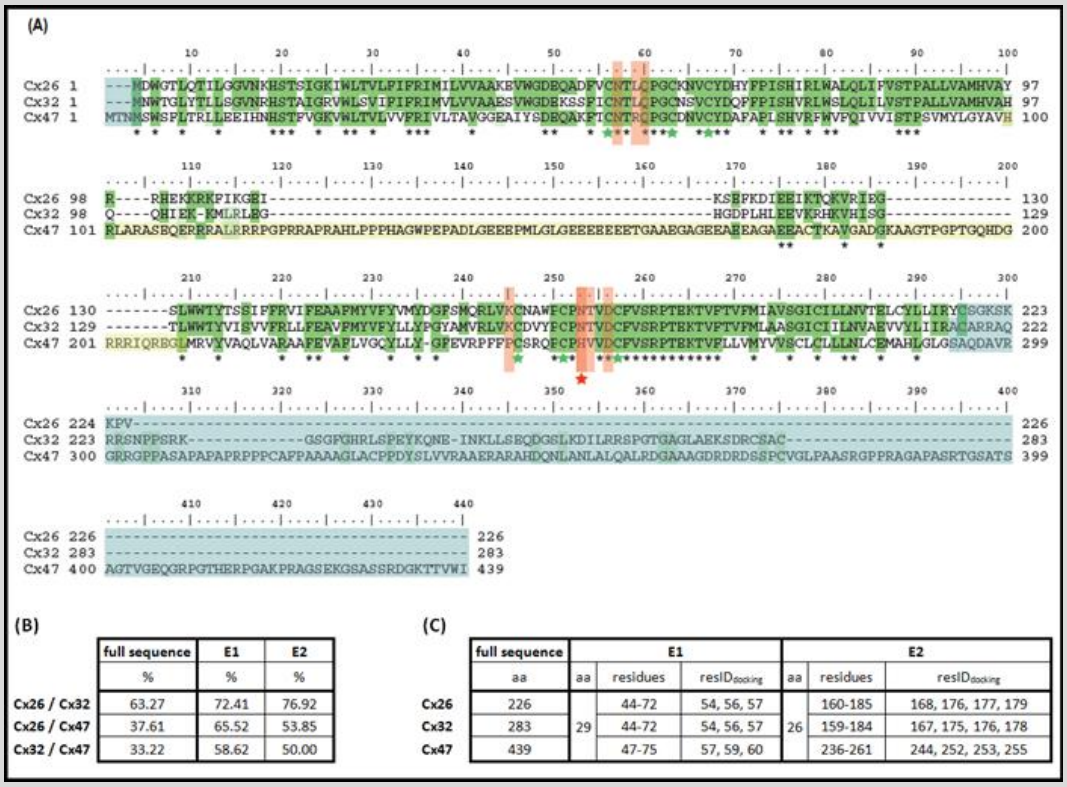
Although the experimental techniques for obtaining protein structures have improved significantly, they still cannot provide a valid structural elucidation for the huge amount of sequences that are known and are constantly being found in genomic analyses. Theoretically, a structure can be predicted from the sequence alone with *ab initio* modelling. Nonetheless, this is considered as a very challenging and time consuming process which successes are limited to small protein domains. At this point, one computational solution is homology (comparative) modelling. The use of this method is based on the observation that two proteins belonging to the same family and sharing similar amino acid sequences, will have similar three-dimensional structures. In other words, the problem is reduced to find a modelling template. The steps are usually the following: template identification, amino acid sequence alignment, alignment correction, backbone generation, generation of loops, side chain generation and optimization, *ab initio* loop building, overall model optimisation and model verification. The accuracy of this method depends on the sequence similarity and the accuracy of the alignment between the proteins under comparison. At low sequence identity (< 30%), the similarity argument cannot be used. This is a rather popular technique and there are many programs and servers available for homology modelling, automatizing the procedure and facilitating its verification.[62][63]

In the presented work:

Connexin selection and sequence alignment. Human protein sequences for Cx26, Cx32, Cx47 were obtained from the UniProt database[64]: Cx26 (entry P29033, gene GJB2 in the chromosome 13) of 226 amino acids; Cx32 (entry P08034, the gene GJB1 in the chromosome X) of 283 amino acids and Cx47 (entry Q5T442, gene GJC2 in the chromosome 1) of 439 amino acids. The sequence alignment was performed using the program ClustalW2 from the EBI package[65]; sequence identity between connexins are shown in Figure 3-1B.

The sequence alignment allowed identifying residue 176 in Cx26 as residue 175 in Cx32 and residue 252 in Cx47. For connexins in the Group A (Table 1-2) this residue is mostly a histidine, whereas connexins in the Group B have always an Asp in His position.[66] On this basis we qualified Cx47

as a member of the Group A. Cx26 and Cx32 were previously reported as members of the Group B.[23] Fragments with lower homology are found in the cytoplasmic domains CL and CT which do not participate directly in the docking process and thereby have been omitted for the docking models. On the contrary, extracellular loops E1 and E2, considered as the "docking domains", show a very high degree of conservation among all the isoforms (Figure 3-1B).



(B)

	full sequence	E1	E2
	%	%	%
Cx26 / Cx32	63.27	72.41	76.92
Cx26 / Cx47	37.61	65.52	53.85
Cx32 / Cx47	33.22	58.62	50.00

(C)

	full sequence	E1			E2		
	aa	aa	residues	$resID_{docking}$	aa	residues	$resID_{docking}$
Cx26	226	29	44-72	54, 56, 57	26	160-185	168, 176, 177, 179
Cx32	283		44-72	54, 56, 57		159-184	167, 175, 176, 178
Cx47	439		47-75	57, 59, 60		236-261	244, 252, 253, 255

Figure 3-1. (A) Sequence alignment of Cx26, Cx32 and Cx47: Green – identical residues between connexins, green stars – cysteine residues stabilizing E1 and E2 loops, yellow – Cx47 cytoplasmic loop exchanged during the modelling by the corresponding Cx32 region, blue – regions omitted while modelling (N- and C-termini), red – residues specified in the docking procedure, red star – the mutation position. (B) Connexin sequences identity with emphasis on docking region (loops E1, E2). (C) Residues specified and considered for the docking process.

Connexin modelling. The Cx32 and Cx47 structures –wild type and mutants – were obtained by homology modelling based on the high resolution structure of Cx26 (2ZW3)[24]. The mutations under consideration were incorporated into the sequences using BioEdit software[67], prior to homology modelling. The cytoplasmic loop of Cx47 (His100-Leu209) was found to be highly distorted in the homology model and it was replaced by the corresponding fragment of Cx32 (His97-Leu131). This substitution did not affect the docking interface and facilitated later hemichannel construction. Different homology servers were tested and the results were cross-validated using the SAVES server[68]. The SAVES server employs different algorithms (e.g. PROCHECK and ERRAT's) for checking protein structures during and after model refinement: PROCHECK performs a Ramachandran plot analysis. For Cx32 and Cx32_Asn175Asp: 85.6 %, 11.3 %, 1.5 % and 1.5 %, of residues were in most favoured, additional allowed, generously allowed and disallowed regions respectively. For Cx47 and Cx47_His252Asp the equivalent figures were 82.7 %, 14.7 %, 1.6 %, 1 %. ERRAT's determines the Overall Quality Factor expressed as the percentage of the protein for which the calculated error value falls below the 95 % rejection limit, which was

91.209 % for Cx32 and Cx32_Asn175Asp, 75.691 % for Cx47, and 75.824 % for Cx47_His252Asp. After validation, the best scored structures (the CPHmodels-3.2 server results) were further processed.[69,70] Eventual clashes at this and subsequent stages were removed manually using the Mutagenesis option of PyMol[71]. Missing residues of Cx26 (mostly in the cytoplasmic loop region of the protein) and its mutants (Cx26_Asn176Asp, Cx26_Asn54His, Cx26_Asn54Ile, Cx26_Asn54Lys, Cx26_Asn54Ser), were also modelled using the same protocol. Any changes on the text files of the structures (renaming chains, renumbering residues etc.) were performed at the Bioinformatics Software web site[72].

Hemichannel construction. Hemichannels were generated by superposing the target chain on a Cx26 wild type (2ZW3) using the Magic Fit routine from the program Swiss-PdbViewer[73]. Hydrogens were added using the program VMD[74]. When introducing a mutation to homo- or heteromeric hemichannels, the fourteen possible combinations between chains were considered and subsequently generated (Figure 3-2). Hemichannels were not merged into channels before the docking or simulations.

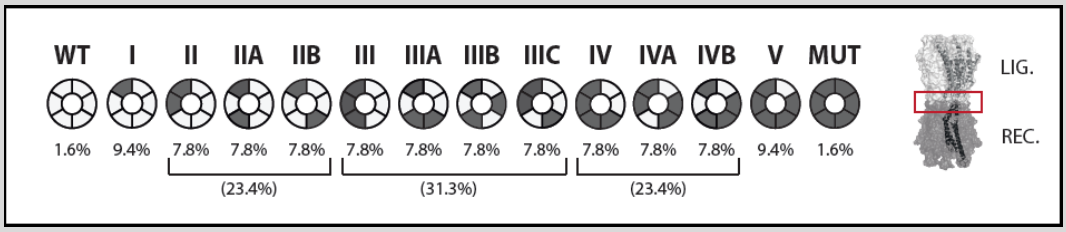


Figure 3-2. Heteromeric hemichannel forms and probability of its occurrence (left), schematic representation of a docked channel (right).

3.1.2 Molecular mechanics (MM) and force fields (FF)

MM refers to the description of the structure and dynamics of molecular systems using the laws of classical mechanics where (i) atoms are described as particles with assigned mass, (ii) bonds between each atom are treated as “springs” with a particular equilibrium distance and force constant, (iii) motions of the atoms are determined by their position, mass and by the forces acting upon them; only the nuclear motions are considered explicitly. Hence, MM is a simplification of the more fundamental quantum mechanics (QM) theory, which treats both the motion of nuclei and electrons. MM allows treatment of much larger systems than QM approaches and is thus particularly suitable for biological macromolecules. A MM potential energy function is used to predict the energy associated with a molecule’s given conformation.[62]

Inextricably linked to the MM issue is a force field (FF) – a collection of atom classifications parameters and analytical expressions for the energy functions that describe the potential energy of a molecular system related with its chemical structure and conformation. Molecular mechanics FFs comply with additivity and transferability principles. They allow expressing the energy of a molecule as a sum of potentials and transferring parameters developed for model chemical groups to similar groups forming a system of interest (a molecule). The potential energy function of a FF (U, Equation 3-1) depends on the atomic positions (r, usually expressed in term of Cartesian coordinates) of all the atoms in the system. It is calculated as a sum of internal terms (bonds, angles

and bond rotations) accounting for atoms linked by covalent bonds (E_{bonded}, Figure 3-3) and a sum of external terms that describe the interaction between atom pairs separated by at least three bonds ($E_{non\text{-}bonded}$). The non-bonded interactions ($E_{non\text{-}bonded}$) are, in particular, long range interactions and depend on the properties of the surrounding medium (charge, presence of ions etc.). Because they involve all the atom pairs, their complexity grows exponentially with the number of atoms thus, causing high calculating costs. To omit the problem, different categories of cutoff techniques neglecting interactions beyond some distance are employed.[62][75][76]

$$U(r) = E_{bonded} + E_{non-bonded}$$

$$E_{bonded} = E_{bond} + E_{angle} + E_{dihedral} + E_{improper} + E_{Urey-Bradley} =$$

$$= \sum_{bond} K_b(b - b_0)^2 + \sum_{angle} K_\theta(\theta - \theta_0)^2 + \sum_{dihedral} K_\phi(1 + \cos(n\phi - \delta)) + \sum_{improper} K_\varphi(\varphi - \varphi_0)^2 + \sum_{Urey-Bradley} K_{ub}(r_{1,3} - r_{1,3;0})^2$$

$$E_{non-bonded} = E_{van\ der\ Waals} + E_{electrostatic} =$$

$$= \sum_{van\ der\ Waals} \varepsilon_{ij}\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \sum_{electrostatics} \frac{q_i q_j}{4\pi r_{ij}}$$

Equation 3-1. CHARMM27 force field energy function decomposition. E_{bond}, E_{angle} – sums over all bonds between two atoms and all angles between two consecutive bonds, respectively; the energy is calculated as a function of displacement from the ideal bond and angle values (b_o, θ_o) with given bond and angle strength (force constants k_b, k_θ). $E_{dihedral}$ –dihedral (torsion) angle (ϕ) around the middle bond for all atoms separated by three covalent bonds, expressed as a cosine function with a coefficient of symmetry (n) and a phase shift (δ). $E_{improper}$ –dihedral term accounting for "out of plane" bending used to maintain chirality and planarity. $E_{Urey\text{-}Bradley}$ – cross-term accounting for angle bending of atoms separated by two covalent bonds. $E_{van\ der\ Waals}$ – interactions between two atoms (i,j) accounting for the balance between repulsive, (short distance) and attractive (long distance) forces modelled with the Lennard-Jones 6-12 potential; ε – energy at the well depth, R_{min} – distance at the potential minimum. $E_{electrostatics}$ – electrostatic interactions between all pairs of atoms represented by a Coulomb potential; D – effective dielectric constant for the medium, r_{ij} – distance between two atoms of partial charges q_i,q_j.

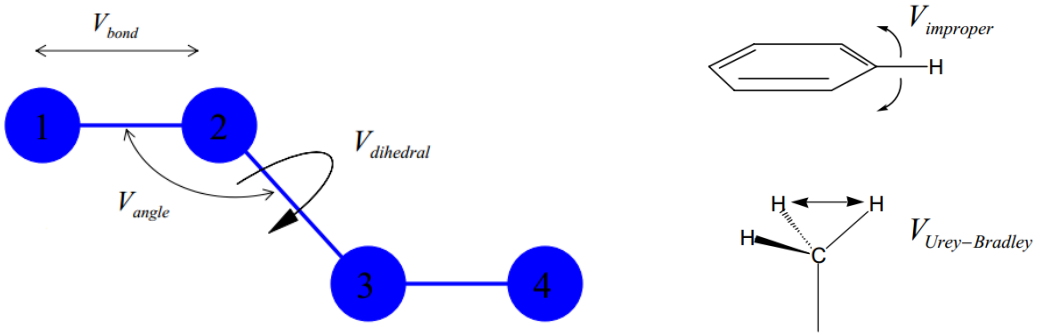


Figure 3-3. Representation of internal terms (E_{bonded}) included in potential energy function (U).[76]

Most FF dedicated to biomolecules (CHARMM, AMBER, GROMOS) have a very similar energy function which parameters (improving the function) are derived from experimental and *ab initio* data. It is the most crucial part of any simulation, presenting a good compromise between accuracy (faithful representation of the interaction between atoms) and computational efficiency (a simple mathematical function that can be calculated quickly). MM approaches are widely applied in molecular structure refinement, molecular dynamics (MD) simulations, Monte Carlo simulations, and ligand-docking simulations.There is no "universal" FF, yet different FFs are specifically designed for different purposes.[62][75][77]

In the presented work:

CHARMM27 FF [78,79] was used to represent proteins and the solvent explicitly.

3.1.3 Molecular dynamics (MD) simulations

The static view of a biomolecule, as obtained from e.g. X-ray or NMR, while extremely valuable only provides an average, frozen view of a system. Thus insufficient for understanding a wide range of biological activity. MD simulations allow calculation of the time dependent behaviour of a molecular system, providing detailed information on its fluctuations and conformational changes. In the first step of a simulation initial positions (those obtained from X-ray or NMR) and velocities are assigned to all particles of the system and Newton's laws are applied to propagate the system's motion through time.[80][81]
Newton's second law (Equation 3-2) states that the sum of all forces acting on an atom results in its acceleration; the acceleration is the second derivative of the position with respect to time and thereby the rate of change of the velocity, which in turn is the rate of change of the position.

$$\overrightarrow{F_i} = m_i\overrightarrow{a_i} = m\frac{d\overrightarrow{v_i}}{dt} = m\frac{d^2r_i}{dt^2}$$

Equation 3-2. Newton's second law of motion. F – sum of forces acting on an atom (i); m – mass, a – acceleration, v – velocity, r – position of an atom (i); t – time.

The force that can also be expressed as the gradient of the potential energy (Equation 3-3),

$$\overrightarrow{F_i} = -\frac{dU(r)}{dr_i}$$

Equation 3-3. Force as gradient of potential energy. U – potential energy of a system, r – system coordinates.

leads to the expression underlying MD method (Equation 3-4).

$$m\frac{d\overrightarrow{v_i}}{dt} - \frac{dU(R)}{d\overrightarrow{r_i}}$$

Equation 3-4. Newton's laws of motion as the basis of molecular dynamics simulation; evaluated form.

Initial positions of atoms – as mentioned before – are assigned by reading in the starting structure (PDB file). Initial velocity assignment (Maxwell-Boltzman distribution) is based on the equipartition theorem that the total kinetic energy of a system is shared equally amongst all energetically accessible degrees of freedom of a system and the following relation is hold in thermal equilibrium (Equation 3-5):

$$E_K(t) = \sum_{i=1}^{N} \sum_{a=x,y,z} \frac{m_i v_{i,a}^2(t)}{2} = \frac{3N}{2} k_B T(t)$$

Equation 3-5. Assignment of velocities complying with Maxwel-Boltzman distribution. E_K – total kinetic energy, N – total number of degree of freedom, k_B – Boltzman constant, T – temperature.

By scaling all velocities we can adjust the instantaneous temperature T(t) to match the desired experimental temperature.

The potential energy (U(r)) is represented through the FF (described above), accounting for the interaction of that atom with the rest of the system. Because the van der Waals and electrostatic interactions exist between every non-bonded pair of atoms, computing the long-range interaction exactly is unfeasible. To perform this computation, the non-bonded forces are spatially truncated at a user-specified cut-off distance. In addition to the intrinsic potential described by the FF, there is a possibility to apply external forces to the system, guiding it into configurations of interest (available in the steered and interactive MD).[75][80]

Due to the complexity of the potential energy function when considering all the atoms, there is no analytical solution to the equation of motion. Instead, an approximate solution is obtained via numerical algorithms that allow their integration, among them – Verlet algorithms (Equation 3-6). The method obtains the position and velocity at the next time step (*r(t+Δt), v(t+Δt)* from the current one (r(t), v(t)), assuming the force is already computed:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{1}{2}a(t)(\Delta t)^2$$

$$v(t + \Delta t) = v(t) + \frac{a(t) + a(t + \Delta t)}{2}\Delta t$$

Equation 3-6. Integration of the equations of motion; Verlet algorithm.

The choice of correct integration time step (dt, Δt) allows for a stable and accurate simulation; the motions of biomolecules involve timescale ranging from 10^{-14} s for an atom bond stretch, to 1 s for the folding of a protein. It is usually chosen an order of magnitude lower than the period of the fastest motion – the vibration of the C-H bond – around 1 fs. This implies millions of calculation steps needed to reach the slow motion that are of biological interest. To overcome this computational problem without significant loss of accuracy, the fastest motion constraints and the multiple time stepping (MTS) scheme were implemented. Fixing some bonds or angles removes the highest frequency motions from the system. Hydrogens, the lightest atoms, are the typical target of this approach – restraining bonds between hydrogens and other atoms permits to increase the time step by a factor of two. MTS scheme – based on the principle that forces between distant atoms exhibit slower fluctuations than forces between close atoms – allows to use three different time

stepping, where the short, intermediate and long range interactions are calculated less often in this order.[62][75]

Moreover, the Newtonian equations of motion has to be modified "mildly" to generate the correct ensemble distribution for the specified temperature and/or pressure (Equation 3-7). After a mentioned random distribution used to set up initial velocities, later on it is ensured by using different thermostats. An example is the Langevin dynamics. The method uses the Langevin equation for a single particle:

$$m_i\vec{v_i} = -\frac{dU(R)}{dr_i} - \gamma_i m_i \vec{r_i} + A_i(t),$$

Equation 3-7. Langevin thermostat adapted to the Newton's equation of motion; γ – fraction coefficient, A – random force.

where the ordinary force the particle experiences (first term) is accompanied by two additional terms used to maintain particle kinetic energy to keep system temperature, for instance, at a constant value. The second term represents a frictional damping that is applied to the particle with frictional coefficient γ. The third term represents random forces which act on the particle as a result of solvent interaction.[62][75]

The initial geometry assigned to the biomolecule under study may not correspond to the actual minima in the FF being used (e.g. overlapping atoms or too close/far distances, often occurring in the coordinate files of deposited structures). Therefore, a simulation is usually preceded by an potential energy minimization to reach the stable and most probable state of a system. The energy landscape of a biomolecule possesses an enormous number of minima, or conformational sub-states and the goal is to find the local energy minimum, i.e., the bottom of the energy well occupied by the initial conformation, even the energy at this local minimum may be much higher than the energy of the global minimum. Physically, energy minimization corresponds to an instantaneous freezing of the system; a static structure in which no atom feels a net force corresponds to a temperature of 0 K.[62][81]

The most common procedure to compare and evaluate results from a simulation (chosen frames of a simulation) with the available structural data, is to calculate the Root Mean Square Deviation (RMSD) between them. This measure estimates the mean square distance between the equivalent atoms of the two superimposed structures:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2}$$

Equation 3-8. Root Mean Square Deviation, where d is the distance between N pairs of equivalent atoms.

An RMSD value is expressed in length units (Å); the smaller value of the comparison, the more similar the structures.

In the presented work:

All MD simulations were performed with the NAMD2[75,82] program along with the CHARMM27 force field[78,79].

Simulations were carried out in the constant volume-constant temperature (NVT) ensemble with the Leapfrog Verlet integrator employing an integration time step of 2 fs. A protein was solvated in a water sphere and the system was neutralized with NaCl. The centre coordinates and the radius of the sphere defined the spherical harmonic boundary conditions with force constant of 10 kcal·mol-1·Å-1 keeping the system together. To maintain the constant temperature of 300 K the Langevin dynamics with 5 ps-1 coupling coefficient was applied to non-hydrogen atoms. Covalent bonds between hydrogen atoms and heavy atoms were constrained by means of the RATTLE algorithm. A switching function was used to truncate the non-bonded interactions (Van der Waals and electrostatics) in the cut-off distance of 12 Å.
The conjugated gradient method of minimization was used to relax the structure before the production for 5000 of force evaluations. MD runs were performed for 2, 5 or 10 ns (the minimization not included). Coordinates to trajectory file were written every 50 or 100 steps and energy value were outputted every 10 steps.
VMD was used for the system preparation and the result analysis of energy, RMSD and hydrogen bonds occupancy.

Protein structures used for simulations:

- Cx47 hemichannel based on the structure with the PDB entry code 2ZW3[24] and prepared according to the protocol demonstrated in the 3.1.1.1 section
- UROIIIS based on the structure with the PDB entry code 1JR2[43]

3.1.4 Molecular docking

Automated docking is widely used to predict biomolecular complexes in structure/function analysis and in molecular design. As in the above case of FFs, there is no single method or program for docking. There are many effective algorithms available. Despite diversity, all docking simulations require a search method for exploring the conformational space available to the system and a method to evaluate the energetics of each conformation, combining it with the desire of reasonable computational effort.

3.1.4.1 ClusPro: protein-protein docking

The goal of protein-protein docking is to take the three-dimensional coordinates of two unbound proteins which are known to interact, and to derive a model for the bound structure, i. e. a complex in atomic details.
ClusPro 2.0 [83–86] is an automated, web-based program for the computational docking of proteins. It performs three steps as follows: (i) rigid-body docking, (ii) clustering of the generated structures and (iii) refinement of the selected structures. The output is a short list of putative complexes ranked according to their clustering properties.
Docking: Most of the current docking methods start with rigid body docking to generate a large number of docked conformations; the ClusPro's core is a Fast Fourier Transform-based (FFT) docking program called PIPER.[84] FFT docking algorithms perform exhaustive evaluation of

simplified energy functions in discretized 6D space of mutual orientations of the protein partners. The space is represented as a grid of displacements of the ligand centre of mass with respect to the receptor's centre of mass; the receptor's centre of mass is fixed at the origin of the coordinate system. ClusPro samples 70 000 rotations of the ligand, which approximately corresponds to sampling at every five degrees in the space of Euler angles. Increasing the number of rotations generally improved the results. Thus, the number of points is chosen as a compromise between performance (representing protein structures with sufficient details) and computational efficiency. The algorithm selects the smallest grid that can accommodate the two proteins, and is efficient for the Fast Fourier Transform. The energy function (scoring function) is given as the sum of terms representing shape complementarity, electrostatic, and desolvation contributions (Equation 3-9):

$$U = E_{shape} + \omega_2 E_{elec} + \omega_3 E_{pair}$$

$$E_{shape} = E_{attr} + \omega_1 E_{rep}$$

$$E_{elec} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \frac{q_i q_j}{\left(r_{ij}^2 + D^2 exp\left(\frac{-r_{ij}^2}{4D^2} \right) \right)^{\frac{1}{2}}}$$

$$E_{pair} = \sum_{i=1}^{N_R} \sum_{j=1}^{N_L} \varepsilon_{ij}$$

Equation 3-9. ClusPro scoring function.

E_{shape} accounts for both attractive (E_{attr}) and repulsive (E_{rep}) interactions which are contributions to the van der Waals interaction energy. The repulsive interactions are cut off at the van der Waals radius plus 2 Å. The attractive part has the cut off radius (6 Å). Users have an option to specify both of them while submitting a job. The difference between the two is that "Repulsion" absolutely prohibits binding to the regions blocked, whereas "Attraction" does not exclusively attract to the selected residues and putative ligands may still be found elsewhere. The electrostatic term, E_{elec} is given by a simplified generalized Born-type expression. E_{pair} represents the desolvation contributions (two atom-level potentials). E_{pair} has been parameterized on a set of complexes that included a substantial number of enzyme–inhibitor pairs and multisubunit proteins, and hence the resulting potential assumes good shape and electrostatic complementarity. ε_{ij} is the energy contribution by a pair of interacting atoms a_i and a_j, and the sum is taken over all pairs of atoms that are closer to each other than a cut-off distance D of 6.5 Å ($d < r_{ij} < D$; whereas $\varepsilon_{ij} = 0$ if $r_{ij} > D$). N_R and N_L denote the numbers of atoms in the receptor and the ligand, respectively. The parameters that differ between different types of complexes are the coefficients ω_1, ω_2, and ω_3 of the energy terms in the expression. Optimized using a small subset of benchmark proteins from the Protein Data Bank, they specify the weights of the corresponding terms ($\omega_1 = 0.40_{(1)(2)(3)(4)}$; $\omega_2 = 0.40_{(1)(2)(3)}, 0.10_{(4)}$; $\omega_3 = 600_{(1)(3)(4)}, 1200_{(2)}$; $\omega_4 = 1_{(1)(2)}, 2_{(3)}, 0_{(4)}$). They are optimally selected for different types of docking problems and are independent of the type of the proteins to be docked.

Clustering (Scoring by pairwise RMSD): The biophysical meaning of clustering is to isolate highly populated low energy basins of the energy landscape – the largest clusters represent the most likely models of the complex, i.e. the closest to the native structures. PIPER does not do filtering, instead it clusters the top 1000 energetically favourable structures using pairwise RMSD as the distance measure. For each docked conformation, the residues of the moving molecule (designated as the ligand) that have at least one atom within 10 Å of any atom of the still molecule (designated as the receptor) are identified and listed. Then, the distance between the Cα of each of those residues and the Cα of the corresponding residues in all other ligands is calculated and subsequently stored into a matrix. Clusters are then formed by selecting the ligand that has the most number of neighbours below a (default) cluster radius of 9 Å. Each member within the cluster is then eliminated from the matrix to avoid overlaps between clusters. This is repeated until at least 30 clusters are formed. The ligand with the most neighbours is the cluster centre, and is considered as the representative structure for the cluster.

Refinement (of the selected docked structures): After clustering, the ranked representative complexes are subjected to a straightforward (300 step and fixed backbone) van der Waals minimization using the CHARMM potential to remove potential side chain clashes. Minimizing the Charmm energy is the only refinement. While minimization generally removes potential steric clashes, it does not substantially change the conformation of the complexes.

The server generates four types of models using the scoring schemes called (i) “Balanced”, (ii) “Electrostatic-favoured”, (iii) “Hydrophobic-favoured”, and (iv) “Van der Waals+Electrostatics”. It is believed that the near native structure will be among the top listed models.

In the presented work:

Hemichannel docking and channel refinement. All docking procedures were run on ClusPro server[83–86]. Residues proven to be active in the hemichannel interactions in Cx26[24] and the homolog residues in Cx32 and Cx47 were selected to drive the complex formation and were taken into consideration for the calculation of the complex binding energy (Figure 3-1C). Specifically, for each chain out of six in Cx26: Asn(Asp)54, Leu56, Gln57, Lys168, Asn/Asp176, Thr177, Asp179, for Cx32: Asn54, Leu56, Gln57, Lys167, Asn/Asp175, Thr176, Asp178 were considered and for Cx47: Asn57, Arg59, Gln60, Pro244, His/Asp252, Val253, Asp255 were considered. The docking output produces a set of complexes ranked according to their clustering properties. For each cluster, the lowest energy member (according to the Balanced mode energy criterion) was chosen as the representative structure. In all cases, the electrostatic-favoured criterion produced very similar results, suggesting a strong electrostatic character for the hemichannel interactions. The rigidity of the receptor in the ClusPro method was taken into account; swapped conformation energies were calculated for all the channels and the average value were taken to the further analysis.

Analysis of structural and stability data. For every set of channel isoforms, docking energies for all complexes were divided by their molecular weight. The normalized dataset of energies are arranged in tables for comparison (for example see Figure 4-4) where, in each case, a lower (higher) energy is equivalent to a stronger (weaker) complex interaction. Energies are coloured according to the interaction strength (green/red for a strong/weak interaction). The normalized numbers for homomeric homotypic (WT/WT, MUT/MUT) and homomeric heterotypic channels (WT/MUT)

were used for the direct comparison with the experimental GJ yields when available; WT/WT and MUT/MUT served as reporters for the GJ stability (100 % and 0 %, respectively). Those cases correspond to the experiment when coupled cells express one-type connexin, wild type or the mutant. To characterize the experiment when one or both of coupled cells were co-injected wild type and the mutant connexin types and the heteromeric heterotypic channels were created, the probability was included into the calculations – in an additional step the energies were multiplied by the chance of a specific hemichannel form occurrence and added together to represent the "mix" system. Energy calculation and data analysis were done with the help of in-house built scripts in AWK and Excel.

Structural analysis of the docked systems focused on the atomic contacts and hydrogen bond network at the interface between hemichannels. Contact maps were created using the CMA tool from the SPACE package.[87] A contact is defined as two atoms with Van der Waals surfaces closer than 1.0 Å. The following Van der Waals radii were used: C: 1.8 Å; O: 1.4 Å; N: 1.7 Å; S: 2.0 Å to generate a two dimensional plot. Little squares indicate that the corresponding residues on the two axes make at least one atomic contact in the PDB file.[88] Hydrogen bonds were investigated by using PyMol with Python scripts (list_hbonds.py[89])using the cut-off of 3.2 Å, the angle of 55° and the acceptor atoms of N or O. Molecular representations were created using PyMol and VMD.

3.1.4.2 AutoDock: protein-ligand docking

To predict bound conformation in a case of protein-ligand complex one may use e.g. MD method with its highly sophisticated FF and full atomic simulation, searching only a small portion of the conformational space and yielding accurate estimates of the energetics of molecular processes. However, these method is too computationally demanding for a docking when (i) no a priori knowledge of the binding site or its location on the macromolecule is available, (ii) an experiment implies compound library screening. Computational docking is typically performed by employing a simpler FF and exploring a wider region of conformational space. AutoDock combines such FF with a Lamarckian Genetic Algorithm (LGA), providing fast prediction of bound conformations with its free energies of association.[90]

In order to allow searching of the large conformational space available to a ligand around a protein, AutoDock uses a grid-based method that allows rapid evaluation of the binding energy of trial conformations. In this method, the target protein is embedded in a grid. Then, a probe atom is sequentially placed at each grid point, the interaction energy between the probe and the target is computed, and the value is stored in the grid (of energies).[90]

The primary method for conformational searching is LGA. A population of trial conformations is created first, and then in successive generations these individuals mutate, exchange conformational parameters, and compete in a manner analogous to biological evolution, ultimately selecting individuals with lowest binding energy.[90][91]

To estimate the energetics of the process of binding of molecules in a water environment AutoDock4 uses a semi-empirical free energy FF which is based on a comprehensive thermodynamic model that allows incorporation of (i) intramolecular energies into the predicted free energy of binding which is performed by evaluating energies for both the bound and unbound

states, (ii) a charge-based desolvation model that uses a typical set of atom types, including the favourable energetics of desolvating carbon atoms as well as the unfavourable energetics of desolvating polar and charged atoms; this differs from a traditional molecular mechanics FF, which typically uses explicit water molecules to evaluate solvation contributions, (iii) improved model of directionality in hydrogen bonds, now predicting the alignment of groups with multiple hydrogen bonds such as DNA bases. The method has been calibrated on a set of 188 diverse protein-ligand complexes of known structure and binding energy, showing a standard error of about 2–3 kcal/mol in prediction of binding free energy in cross-validation studies.[90][92]

The free energy of binding (ΔG) is estimated to be equal to the difference between the energy of the ligand and the protein in a separated unbound states and the energy of the ligand–protein complex. The FF includes six pair-wise atomic terms (U), including evaluations for dispersion/repulsion, hydrogen bonding, electrostatics, and desolvation and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta G = (U_{bound}^{L-L} - U_{unbound}^{L-L}) + (U_{bound}^{P-P} - U_{unbound}^{P-P}) + (U_{bound}^{P-L} - U_{unbound}^{P-L} + \Delta S_{conf})$$

Equation 3-10. Free energy of binding (ΔG). L – ligand, P – protein. First two terms – intramolecular energies for the bound and unbound states of the ligand, following two – intramolecular energies for the bound and unbound states of the protein; in the case of rigid docking motions in the protein are not allowed so the bound state of the protein (U_{bound}^{P-P}) is identical with the unbound state ($U_{unbound}^{P-P}$), and the difference in their intramolecular energy is zero. Third parentheses – changes in intermolecular energy between the bound and unbound states; it is assumed that the two molecules are sufficiently distant from one another in the unbound state that $U_{unbound}^{P-L}$ is zero.

$$U = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) + W_{hbound} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon r_{ij} r_{ij}} + W_{sol}(L_i V_j + L_j V_i) e^{\left(-r_{ij}^2 / 2\sigma^2\right)}$$

Equation 3-11. Pair-wise evaluations for (ΔG). W – empirically determined weighting factors, optimized to calibrate the empirical free energy based on a set of experimentally characterized complexes. First term – typical 6/12 potential for dispersion/repulsion interactions. A, B - parameters taken from the Amber FF. Second term – a directional H-bond term based on a 10/12 potential. C, D – parameters assigned to give a maximal well depth of 5 kcal/mol at 1.9 Å for O-H and N-H, and a depth of 1 kcal/mol at 2.5 Å for S-H. E – directionality of the hydrogen bond interaction, dependent on the angle (t) away from ideal bonding geometry. Third term – electrostatic interactions. F_{final} term – desolvation potential based on the volume (V) of the atoms surrounding a given atom, weighted by a solvation parameter (L) and an exponential term based on the distance. σ – distance weighting factor, set to 3.5 Å.

$$\Delta S_{conf} = W_{conf} N_{tors}$$

Equation 3-12. Loss of torsional entropy upon binding (ΔS_{conf}) - directly proportional to the number of rotatable bonds in the molecule (N_{tors}).

Ligand flexibility is assigned in several steps. First, a root atom is chosen, which will act as the centre of rotation during coordinate transformation in the docking simulation. To find the optimal atom, the number of atoms in each branch are evaluated, and the root atom that minimizes the size of the largest branch is chosen. In some cases, the user may wish to limit the flexibility of the ligand or incorporate limited flexibility in the receptor. The user selects specific sidechains and these are separated from the protein. During the simulation, these are treated explicitly, allowing rotation

around torsional degrees of freedom using the same methods used to explore the conformational space of the flexible ligand. The remaining portion of the protein is represented using the affinity grids described above.[90]

AutoDock is accompanied by AutoDockTools – a graphical user interface facilitating the coordinate preparation and analysis. It allows the user to identify the active site and determine visually the volume of space searched in the docking simulation.[90]

In the presented work:

Protein receptor preparation. Porphobilinogen deaminase structure was obtained from the PDB base[93]; protein code – 3ECR[34]. Residues missing mainly in a loop region of a chain A (Ser60-Ser78, Gly262-Gly263) were modelled using SWISS-MODEL server[94,95]. The crystallographic water molecules were removed, the missing hydrogen atoms were added and the cofactor (DPM) from the crystal structure was used to define the active site.

Ligand preparation. Starting with a two dimensional text representation (SMILES string) of a known ligand for a receptor, and at the same time the reaction substrate – PBG, a list of similar compounds was selected using ChemNProp Web Application (cemnprop.irbbarcelona.org; currently unavailable). Examples with the default value of similarity over 0.17 were chosen. Water solubility was calculated using ALOGPS 2.1 program[96,97]; compounds with a solubility below 1g per 100g of water were considered insoluble and rejected. From the final list of around 600 compounds, those commercially available were subjected to further preparations (~220). The chosen ligands were drawn in SymyxDraw 3.2 (developed by Symyx Technologies, Inc.) or retrieved from the DrugBank database[98,99]. ArgusLab[100] was used for file conversion and energy minimisation. Finally, all the compounds were saved in PDB format for further docking studies.

Docking simulations. Python scripts generated a PDBQT files from the receptor and ligand molecular structures using AutoDockTools[101], adding partial charges and atom types to the PDB format. Lamarckian genetic algorithm methodology was employed for docking simulations implemented in AutoDock 4.2. The standard docking procedure was used for a rigid protein and a flexible ligand whose torsion angles were identified; ten independent runs per ligand. A grid centred on macromolecule of 126, 126, and 126 points in x, y, and z directions was built with a grid spacing of 0.458 Å. The default settings were used for all other parameters.

Analysis of docking results. From ligand clusters those with best (lowest) binding energies and localized within the active centre area were analysed for hydrogen bonding using AutoDockTools and Discovery Studio Visualizer 4.5 [102]. Most promising inhibitors were directed to enzymatic assays. 58 were experimentally tested and 15 best results will be discussed in the 4.2.1.2 section.

Superimposition and visualization of the final inhibitor candidate group were done in Swiss PDB Viewer and PyMol programs. From the estimated free energy of ligand binding ($\Delta G_{binding}$, kcal/mol), the inhibition constant (K_i) for each ligand was compared to the experimental results.

3.2 Biophysical methods

3.2.1 Protein sample production and preparation

To uncover and characterize protein structure, function and dynamics it is necessary to obtain samples with high purity (> 97 %) and in large amounts. To that end, structural biologist have adapted to these requirements expression and purification protocols which use different host organisms to overexpress the macromolecule under study. In general, the sample obtainment procedure for structural biology consists of steps of heterologous protein expression, and protein purification.

In the presented work:

The heme pathway proteins, PBGD and UROIIIS, has been studied practically and theoretically by the author and members of the Oscar Millet group, thereby the experimental protocols for them will be described in details; **the enzymes are obtained and purified by following the method published by Millet O et al.**[45].
The connexin proteins have been studied theoretically by the author; all experimental result have been provided by the group of Dr. Louis Barrio.

3.2.1.1 Protein subcloning and expression

Protein expression methods can be divided into two groups: (i) the traditional one in which the enzyme is purified from its natural source, (ii) the modern recombinant protein enzyme production using molecular biology techniques to obtain the raw material by heterologous expression.[103]
Ability to express proteins in a recombinant form has greatly advanced and, together with the exponential rise in knowledge of gene structures, this enables proteins to be produced in a friendlier host organism, without ever using the natural source material (heterologous expression). Expression of foreign proteins in a host organism (most often *E. coli*) has many advantages: (i) the probability that the expressed enzyme will make up a far larger proportion of the starting extract than occurs in the natural source, (ii) the production of the raw material under controlled conditions at a convenient time, (iii) the ability to modify the polypeptide at the gene level to simplify the purification or crystallization processes. For instance, widely used is to add a 'tag' which can be recognized by an affinity adsorbent. A short stretch of amino acids added either to the N- or C-terminal can be used; typically a six-histidine stretch may be added, allowing selective adsorption on an immobilized metal. Alternatively the tag can include sequences that are recognized by antibodies, and many other systems. If necessary, the tag can be latter removed by specific proteolysis. When the protein is soluble in the host organism, and highly expressed, then it becomes the optima situation for the use of conventional techniques to separate away the interest proteins from the host constitutive material.[103]

In the presented work:

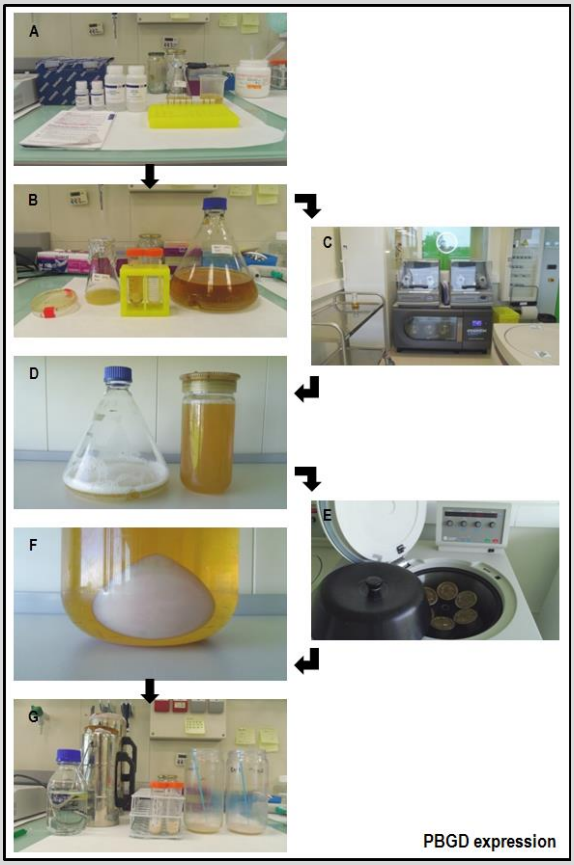


Figure 3-4. PBGD expression steps.

Subcloning and Site-Directed Mutagenesis.

- pHTU3S plasmid containing the gene codifying for human UROIIIS was incorporated to the laboratory provided by Prof. J. Phillips (Utah University).
- Human PBGD corresponding to the housekeeping enzyme isoform (further called PBGD1), required for the UROIIIS specific activity assay, was amplified by PCR reaction using the forward and reverse primers 5′-ACACCATGGCATCTGGTAACGGCAATGCGGCTGCAACGGCG G-3′ and 5′-GGGGAAGCTTTTAATGGGCATCGTTAAGCTGC-3′ and cloned into a pETM-11 expression vector using the *Hin*dIII and *Nco*I restriction sites.
- Truncated human PBGD corresponding to the erythrocyte-specific enzyme isoform (-17aa N-terminal, PBGD2) was amplified by PCR reaction using the forward and reverse primers 5′-TTACATATGAGAGTGATTCGCGTGGGT-3′ and 5′-GGCTGGATCCATTTAATGGGCATCGTTA
- AGGAAT-3′ and cloned into a pET-16b expression vector using the *Nde*I and *BamHI* restriction sites.
- PBGD_R26V (PBGD1_R26V) mutant was obtained based on the housekeeping PBGD isoform in a pETM-11 vector (primers sequence lost) in order to revise the molecular docking result analysis (section 0).

Site-directed mutagenesis of PBGD was carried out employing the commercial QuikChange®II Site-Directed Mutagenesis kit (Stratagene) with custom-made oligonucleotides as PCR reaction primers (Invitrogen). Reactions were set up according to the manufacturer's instructions. The entire length of the genes was sequenced in order to verify the mutations. Mutants were expressed

and purified the same way as the wild-type enzyme and assayed for activity under identical conditions.[45]

Protein expression. Freshly transformed *E. coli* BL21(DE3) cells were used for protein expression. First, cells were plated in petri plates with 50 mg/mL of the corresponding antibiotic (UROIIIS, PBGD2 – ampicillin and PBGD1, PBGD1_R26V – kanamycin) and incubated over night at 37 °C. Next grown at 37 °C for 16 h in small preinocules of 100 mL of LB media (or M9 minimal media for isotopic labelled proteins), separated by centrifugation and resuspended into 1–1.5 L of culture to grow at 37 °C until OD_{600} reached 0.6-0.8. The culture was then induced with 0.5-1 mM isopropyl-β-D-thiogalactoside (IPTG) for 20 h at 16 °C. Cells were harvested by centrifugation, resuspended in 20-30 mL of resuspension buffer per litre of culture and stored at -80 °C prior to purification process. PBGD expression is schematically presented in the Figure 3-4.

Proteins for biophysical and enzymatic activity tests were grown in regular LB media, whereas protein samples required for NMR experiments included an specific **isotopic labelling scheme**: typically ^{15}N, ^{13}C and sometimes ^{2}H. M9 minimal media is a restrictive media in which the sources of nutrients are tightly controlled (composition of M9 minimal media : Na_2HPO_4 – 6 g/L, K_2HPO_4 – 3 g/L, NaCl – 0.5 g/L, $^{(15)}NH_4Cl$ – 1 g/L, $^{(13)}C$-glucose – 2-3 g/L, $MgSO_4$ – 1 mM, $CaCl_2$ – 0.1mM, Biotin – 1 mg, Thiamine hydrochloride – 1 mg, Antibiotic – 0.5 mM, H_2O (D_2O) up to 1L). Ammonium chloride and glucose were used as the sole sources of nitrogen and carbon for protein synthesis. ^{15}N-labelled ammonium chloride. As a result, the addition of $^{15}NH_4Cl$ or ^{13}C-glucose ensures that all nitrogen or carbon atoms in the protein are isotopically labelled.

3.2.1.2 Protein purification

A protein (collected as a cell pellet after the production) has to be thoroughly purified before reaching the following stage of structural characterization. Often more than one purification step is necessary to reach the desired purity. The key to success and efficiency is to select the most appropriate techniques, optimise their performance and combine them in a logical way to maximise yield and minimise the number of steps required, as each step will imply a reduction in the protein yield. Biomolecules are purified using chromatography techniques that separate them according to a different sensitivity based in their specific properties (Figure 3-5). The process can be monitored using UV absorption spectroscopy and/or electrophoresis techniques, commonly SDS-PAGE.[104]

Property	Technique
Size	Gel filtration (GF), also called size exclusion
Charge	Ion exchange chromatography (IEX)
Hydrophobicity	Hydrophobic interaction chromatography (HIC) Reversed phase chromatography (RPC)
Biorecognition (ligand specificity)	Affinity chromatography (AC)

Figure 3-5. Chromatography technique types.[105]

In the presented work:

Protein sample preparation. UROIIIS (MW 31.1 kDa) has been purified employing the following protocol: the cell pellet (1 L of culture) is resuspended in buffer A (20 mM sodium

phosphate, 500 mM NaCl, 5 mM imidazole, pH 7.4) with the addition of Protein Inhibitor Cocktail (PIC Sigma, dilution ~1:100); the sample must be kept at cold temperature on an ice bucket at all times to minimize proteolysis. Pressure induced cell lysis is achieved in a cell disruptor at 27 kpsi (alternatively by thorough sonication) and it is followed by ultracentrifugation in a Beckman Ultracentrifuge at > 30-45k rpm for 30-45 min at 4 ºC. Pellet is immediately discarded, whereas supernatant contained the protein mixed with 3 mL of His-tag resin (Ni-NTA; Invitrogen) and eluted with buffer A containing 500 mM imidazole. Protein is further purified by size exclusion chromatography (Superdex 75; GE Healthcare) under isocratic conditions (20 mM Tris, 150 mM NaCl, pH 8.0). PBGD (MW 42.6 kDa) is purified following the same protocol except eluted in 20 mM Tris and 150 mM NaCl at pH 7.0. Samples concentration is next measured by UV spectrophotometry at 280 nm, employing extinction coefficients of 25410 and 15970 M-1 ·cm-1 for UROIIIS and PBGD, respectively and kept at 4°C or flash frozen and stored at -80C. Between 0.5 and 1.2 mg of pure wildtype UROIIIS and 3.8 mg of PBGD were obtained per liter of culture.[45] PBGD purification is schematically presented in the Figure 3-6 with samples in characteristic pink colour.

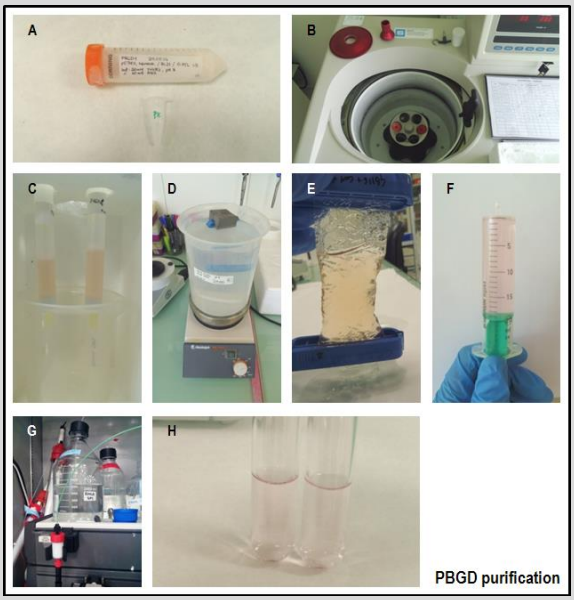


Figure 3-6. PBGD purification steps.

The modified PBGD purification protocol leading to the reaction intermediates separation and applied especially to the structural study procedures is as follow and will be closer discussed in the section 4.2.2.1:

BUFFERS USED: A buffer – 20 mM Tris-HCl pH 8.7-8.8, 40 mM NaCl, B buffer – 20 mM Tris pH 8.7-8.8, 1 M NaCl, "washing" buffer: A buffer + 5 mM imidazole, "eluting" buffer: A buffer + 500 mM imidazole

PROTOCOL:

- The cell pellet (0.75 L of culture) is resuspended in buffer A with the addition of Protein Inhibitor Coctail (PIC Sigma, dilution ~1:100).
- Cell are lysed in a cell disruptor at 27 kpsi (preferably) or by thorough sonication.
- Cell lysate is ultracentrifugated at 30-45k rpm for 30-45 min at 4 ºC. Pellet is discarded, whereas supernatant contained the protein mixed with 3 mL of His-tag resin (Ni-column,

beforehand washed with the "washing" buffer) and left for 1-2h of incubation (4 ºC, if possible – being gently shaked/rotated).

- After re-rinsing column with "washing" buffer for unbound proteins' removal, sample is eluted with "eluting" buffer and left for O/N dialysis. Dialysis serves for (i) imidazole dilution, (ii) His-tag cutting.
- Sample is repeatedly loaded to the Ni-column to separate protein fractions W/ and W/O His-tag. In case of protein precipitation short centrifuging may be needed before loading.
- Sample that has been washed out (W/O His-tag) is subjected to size exclusion chromatography in buffer A in a Superdex 75 column (*HiLoad 26 600)*.
- Fraction containing pure PBGD (Figure 4-23A) is then loaded to anion exchange chromatography column, HiTrap Q HP 5ml or MonoQ 5/50 GL. The resulting enzyme-intermediate complexes are eluted with a linear gradient of buffer B; 0-30% in 35 CV or 0-20% in 50CV for respectively, Q or MonoQ column (Figure 4-23B).
- Separated complexes, which concentration has been measured by UV-vis spectrometry at 280 nm, can be used immediately or flash frozen and stored in -80 ºC.

3.2.2 Protein characterization

Among the experimental methods for characterizing macromolecules and their interactions we can distinguish those focusing on the function and on the structure.

3.2.2.1 Functional properties

3.2.2.1.1 Enzymatic assays

Enzyme activity is measured *in vitro* under well-defined conditions that often do not closely resemble the functional ones found *in vivo*. Nevertheless, with a complete study of the parameters that affect enzyme activity it should be possible to extrapolate to the activity expected to be occurring *in vivo*. For consistent and reproducible results, the parameters of optimum pH, 'saturating' substrate concentrations, temperature, buffer type and ionic strength must be carefully controlled. This will define how much enzyme is present in a sample compared with others, but generally will not indicate the flux of metabolites through that enzyme *in vivo*. The traditional enzyme has a hyperbolic response to substrate concentration, according to the Michaelis–Menten equation (Equation 1-4).[103]

Enzymes are active only within a limited (yet relatively broad) range of pH. For standard enzyme assays, a value of pH is chosen that is close to the optimum (at which the maximum activity (the highest value for V_{max}) is attained), unless some other component of the assay mixture cannot operate at that pH. Temperature affects enzyme activity in much the same way as it affects other chemical reactions. Rates increase by between 4 % and 8 % per degree, although at high temperatures denaturation of the enzyme protein decreases product formation. Thus, it is important when carrying out an enzyme assay to ensure that the temperature remains constant. The activity can also be affected by the nature of the buffer used, so alternative buffers for a given

pH should be tested. Each enzyme responds also in a unique way to ionic strength I (salt content). Most are inhibited at high osmolarity (> 0.5 mol/L). For most enzymes there is a variation of activity with I, so the value should be fixed. A natural intracellular ionic strength is typically in the range of 0.15 to 0.2 mol/L.[103]

To measure the activity of an enzyme is equivalent to measure how much product is formed over a given time or, in some cases, how much substrate has been consumed. It is the aim of a successful enzyme assay to measure the product formation before it has accumulated sufficiently to affect the initial rate.[103]

There are different types of assays. Stopped assays involve interruption of the reaction after a fixed time, and subsequent measurement of the formed product. Stopping methods include those that denature the enzyme, such as strong acid, alkali or detergent; heat; or treatments with irreversible inhibitors such as heavy metal ions. In some cases the enzyme can be stopped by addition of a complexing agent such as ethylenediaminetetraacetic acid (EDTA), which removes metal ions essential for activity or, in other protocols, chilling on ice or in liquid nitrogen may be sufficient. The alternative to a stopped assay is a continuous one in which the progress of the reaction is followed on the fly. The simplest continuous follows the action of the enzyme by changes in a measurable property: UV absorbance, fluorescence, viscosity, pH, or other biophysical parameter.[103]

Enzyme assays not only monitor the characterization of an enzyme activity but also are sensitive to the inhibition mechanism. One can identify inhibitors that have competitive, non-competitive and uncompetitive behaviour with regard to the substrate at or below K_m. An expression of the potency of a compound is described by the IC50, the half maximal inhibitor concentration and is commonly used as a measure of antagonist drug potency in pharmacological research. More advanced cell based assays of biochemical activity and applicable to compound libraries are utilized to fasten the identification of potential inhibitors or to identify promising molecules in throughput screening.[106,107]

In the presented work:

In vitro assays. Enzymatic assay: The assay is based on the method described by Jordan[108] with some modifications[60]. 10 µL of PBGD at a concentration of 150 µM in 48 µL of buffer (20 mM Tris-HCl pH 8.0, NaCl 150 mM) is preheated at 37 °C. for two minutes. If the assay is performed in the presence of an inhibitor, the latter is added in dissolved form into the buffer. 25 µL of the PBG substrate are then added, diluted to different concentrations (0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 10.0, 13.0, 15.0 mM) in the aforementioned buffer and are incubated at 37 °C. for five minutes. A control assay containing a buffer without substrate (PBG) in the same proportions is performed at the same time. The enzymatic reaction is stopped by means of freezing the samples in liquid nitrogen. Then the mixture is incubated again with 6 µL of 25 mM UROIIIS at 37 °C. for 30 minutes to transform the pre-uroporphyrinogen into URO III. The reaction is again stopped by means of freezing in liquid nitrogen. Then the formed CO2 and uroporphyrinogens are oxidized to uroporphyrins by means of adding 25 µL of I2/KI solution and incubation for 10 minutes in ice and darkness. The iodine excess is then bleached by adding 25 µL of saturated sodium metabisulfite solution. 250 µL of 10 % trichloroacetic acid (TCA) are added to precipitate

the protein and the samples are centrifuged at 13 000 r.p.m. for 5 minutes. Finally, different dilutions of the supernatant in 0.1 M HCl are prepared to measure the change of absorbance at 405 nm with respect to the control sample.
Maximal reaction rate, Michaelis constant or inhibition constants – if calculated – have been calculated from the global data fitting of all values in presence of inhibitor according to the established methodologies[109].

3.2.2.1.2 Patch clamp

Characterization of the protein activity becomes more challenging when addressing complex systems which work in an environment difficult to mimic *in vitro*, e.g. ion channels. The physiology of ion channels – in general, excitability of membranes and cells, like neurons, cardiomyocytes or muscle fibres - has always been a major topic of interest, especially in neuroscience. An indirect method of determining the channel activity is the patch-clamp technique, which has demonstrated to be an optimal solution. It allows high-resolution recordings of the current not only of whole cells, but also of excised cellular patches. Even single-channel opening events can be investigated, which is remarkable since a single ion channel conducts around 10 million ions per second, yet the current is only a few pico amperes. The general principles of the method are shown in the Figure 3-7.[110]

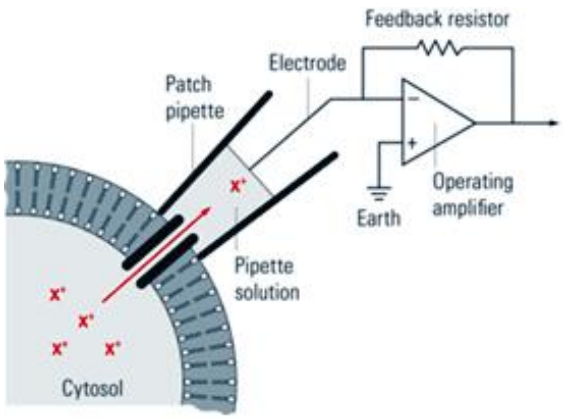


Figure 3-7. General principles of patch-clamp recordings. A glass pipette containing electrolyte solution is tightly sealed onto the cell membrane and thus isolates a membrane patch electrically. Currents fluxing through the channels in this patch hence flow into the pipette and can be recorded by an electrode that is connected to a highly sensitive differential amplifier. In the voltage-clamp configuration, a current is injected into the cell via a negative feedback loop to compensate changes in membrane potential. Recording this current allows conclusions about the membrane conductance.[110]

Depending on the specific measurement, different configurations can be used. In the *cell-attached* mode (and its modification – loose-patch) the membrane patch is left intact. The composition of the cytoplasm is not altered and, therefore, the intracellular environment cannot be controlled. The most commonly used patch-clamp mode is the *whole-cell* mode. To achieve it, the membrane patch is disrupted by briefly applying strong suction. The interior of the pipette becomes continuous with the cytoplasm. This method is used to record the electric potentials and currents from the entire cell. In *whole-cell* measurements the researcher can choose between two configurations: the voltage-clamp mode in which the voltage is kept constant and current is recorded, or the current-clamp mode in which the current is kept constant and changes in the

membrane potential can be observed. Alternatively, it is also possible to record currents only from a small patch instead of the whole cell, i.e. recording single channels. The patch can be oriented in two different directions inside the patch pipette. To achieve the *inside-out* configuration the patch pipette is attached to the cell membrane and is then retracted to break off a patch of membrane. In this case, the cytosolic surface of the membrane is exposed. This is often used to investigate single channel activity with the advantage that the medium that is exposed to the intracellular surface can be modified. If the aim is to study the influence of extracellular cues such as neurotransmitters, the *outside-out* configuration is chosen. In this case the pipette is retracted during the *whole-cell* configuration, causing a rupture and rearrangement of the membrane. In this configuration the extracellular surface is exposed and extracellular cues can easily be applied.[110]
Patch-clamp recordings can be combined with live-cell imaging approaches such as Ca^{2+}, pH- or Cl^- imaging, where a Ca^{2+}, pH- or Cl^- -sensitive fluorescent dyes are applied to the cell via the patch pipette. The membrane current and changes in fluorescence are recorded simultaneously. It is also a commonly employed method in medical research, since many diseases are related to a malfunction of definite ion channels. In pharmacological research, automated patch-clamping is used to screen potent substances for ion channel modifications.[110][111]

In the presented work:

Electrophysiology experiments. All patch-clamp experiments, underlying my theoretical connexin studies, have been conducted by the group of Dr. Luis Barrio.
The preparation of Xenopus oocytes and RNA microinjections was performed as described previously.[112] Oocytes were injected with an antisense oligonucleotide (15 ng/oocyte) that blocks endogenous Cx## expression, either alone (controls) or with Cx## mRNAs (0.5–1.0 g/l, 50 nl/oocyte). To allow intercellular channels to form, pairs of oocytes were brought into contact after removing the vitelline membrane Junctional currents between oocyte pairs were measured using the dual voltage-clamp technique. Pipettes for oocyte patches were filled with external ND96 solution (in mM: 96 NaCl, 2 KCl, 1 MgCl2, 1.8 CaCl2 and 5 HEPES, pH 7.5) or standard internal solution (SIS; in mM: 100 KCl, 10 HEPES, 10 EGTA, pH 7.2). For data acquisition and analysis we used a Digidata interface and pClamp (Axon Instruments, Inverurie, Scotland), as well as Sigmaplot software (Jandel Scientific, Corte Madera, CA, USA). The Student's t test was used for statistical analysis.

3.2.2.2 Structure determination

There are three main experimental techniques to determine the structure of a protein – X-ray methods, nuclear magnetic resonance spectroscopy (NMR) and cryo-electron microscopy (cryoER).

3.2.2.2.1 X-ray methods: Small angle X-ray scattering (SAXS) and X-ray Crystallography

The broad definition of X-ray methods covers many techniques based on the scatter, emission and absorption properties of X-radiation. SAXS and X-ray crystallography are fundamentally similar

techniques exploiting coherent X-ray scattering. Coherent radiation (here mainly photon) upon interacting with the attenuating medium does not have enough energy to liberate the electron from its bound state, so no energy transfer occurs. The only change is a change of direction (scatter) of the photon. When X-ray photons collide with atomic electrons, the scattered wavelengths interfere with one another and diffraction occurs.[113][114]

SAXS and X-ray crystallography can share most of the hardware required to generate, prepare, and detect X-rays, schematically shown at the Figure 3-8 A. An elementary difference between them lies in the organization of target molecules during data collection. In solution scattering, which is inherently a contrast method, the signal from all orientations of the target molecules are averaged together. Specifically, derived from the difference in the average electron density of solute molecules of interest and bulk solvent. Solution scattering is continuous and radially symmetric (Figure 3-8 B). In contrast, in X-ray crystallography the molecules are highly organized within a crystal lattice. Diffraction from a crystal lattice (Figure 3-8 C) gives rise to discrete diffraction maxima that are caused by the convolution of the crystal lattice onto the continuous transform due to the atomic positions and provides enormously greater signal. There is a simple relationship (Bragg's law) between the scattering angle, the wavelength of the radiation and the spacing between the planes of atoms. Since the distances between the atomic planes are dependent on the size and distribution of atoms – i.e. the structure of the material can be used for qualitative and quantitative phase identification.[114]

During the very first step of the diffraction data processing of crystal system, unit cell dimensions and crystal orientation are determined. Next, each spot on the image is assigned an index (quoted as three integers: h, k, and l)and its intensities are measured. A scale factor must be allocated so that the intensities of all the images in the data set can be related.[115]

The intensity of any diffracted spot is the result of the diffracted waves incident on the detector at that point. Therefore, the intensity will be determined by the amplitude of those waves and by the phase difference, expressed as an angle, between them. The determination of the amplitudes is mathematically simple. The square root of the intensity is usually taken, with negative values being set to zero, and outputting amplitudes accordingly. The phase angle however, cannot normally be determined directly in the case of protein crystals and so must be found in an indirect way. The two most frequently used methods are isomorphous replacement and molecular replacement.[115]

Isomorphous replacement requires at least two data sets: one native set from the protein crystal alone and at least one derivative set from the protein crystal with attached heavy atoms. When compared, the differences between these data sets results solely from the heavy atoms and, therefore, their positions in the protein molecules can be determined. Computer refinement of the heavy atom parameters is carried out and these parameters can be used as a starting point to determine the protein phase angles.[115]

Molecular replacement is the most rapid and most frequently used when a very closely related protein structure is available. The method involves the crystallographic calculation in reverse: the phases from the known model structure factors are subsequently "borrowed" and applied to the new data set. Hence, there will always be a certain amount of bias towards the model.[115]

The amplitudes and phases allow calculating, using the FFT method, the structure factors and the resulting electron density map will form the three dimensional contours into which the protein structure will be built. The quality of the electron density map may be improved by refinement.[115]

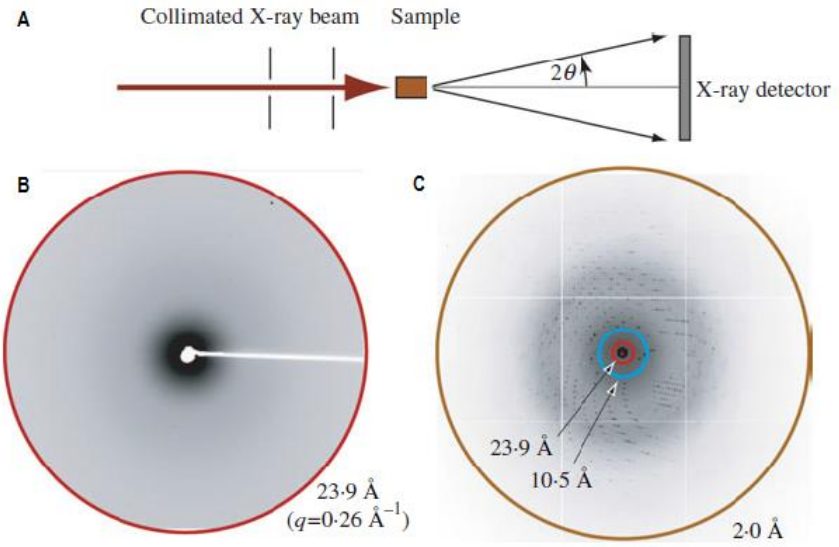


Figure 3-8. X-ray interactions with a sample for SAXS and crystallography. (A) Both SAXS and X-ray crystallography involve placing a sample (orange) into a highly collimated X-ray beam (red) and measuring the scattered X-rays. The angle of any scattered position with the direct beam is 2θ. (B) Example scattering with a maximum resolution of 23.9 Å . (C) Example diffraction at 2.0 Å resolution. The equivalent position of the highest resolution of the SAXS experiment is indicated (red circle). The blue circle indicates the highest resolution achievable (q=0.6 Å) at the used beamline.[114]

Crystallography provides substantially more information content than SAXS, allowing atomic resolution structures (critical for mechanistic analyses) to be determined; however, the requirement of packing in the crystal lattice can lead to molecules whose conformations are inappropriately fixed by non-biologically relevant interactions. SAXS offers complementary low resolution information about macromolecular folding, aggregation, extended conformations, large complexes and conformational states that are too flexible, too large, or too difficult to stabilize as homogeneous samples for these techniques. The resolution range of about 50 A° to 10 A°, but without the size limitations present in NMR and electron microscopy studies.[114]

In the presented work:

SAXS data collection and analysis. Synchrotron SAXS data were collected on beamline BM29 at ESRF (Grenoble, France) with a 2D detector (Pilatus 1M) over an angular range q_{min} = 0.01 $Å^{-1}$ to $_{qmax}$ = 0.5 $Å^{-1}$. X-ray scattering patterns were recorded for PBGD reaction intermediates and PBGD_R26V at the concentration of 1.92 and 6.40 mg/mL (P1), 2.63 and 5.24 mg/mL (P2), 2.10 and 4.26 mg/mL (P3), 1.53 and 3.43 mg/mL (P4), 4.96 and 9,76 mg/mL (MUT) in standard purification buffer (20mM Tris-HCl pH 8.5, 40mM NaCl).
Data collection, processing, and initial analysis were performed using beamline software BsxCuBE. Further analyses were performed with the ATSAS suite[116]. PRIMUS[117] was used for R_g and maximum distance (D_{max}). Fitting of a model structure to the SAXS data was calculated with CRYSOL[118]. In order to generate an ab initio model, ten runs of DAMMIN[119] were performed,

and after averaging and filtering a model with 732 beads was produced. Superposition of the bead model on the relaxed crystallographic PBGD (3ECR) structure was carried out using the program SUPCOMB[120]. The resulting model was visualized using PyMOL.

Crystallization and Structure Determination. Diffraction-quality PBGD crystals for holoenzyme and one of reaction intermediates were obtained by hanging-drop vapour diffusion at 21 °C in a solution containing 0.1 M Na HEPES pH 7.2 (buf.), 0.9 M Na dihyhydrogen phosphate/0.9 M K dihydrogen phosphate (salt) and 0.1 M sodium cacodylate pH 6.5 (buf.), 0.2 M sodium chloride (salt), 2 M ammonium sulphate (salt), respectively. Individual crystals were cryo-protected by a brief soak in well buffer supplemented with 25 % glycerol and flash frozen in liquid nitrogen.

Diffraction data were remotely collected at a wavelength of 0.972422 Å at ID23-1 beamline at the ESRF (Grenoble, France) synchrotron centre. Data were integrated and scaled using XDS program[121]. Molecular replacement was done by the MOLREP program[122] with E. coli PBGD (PDB ID: 3ECR) as the search model. Refinement is currently being performed using COOT[123] and REFMAC[124] of CCP4 suite[125].

Figure 3-9. Equipment used for SAXS (left) and crystal diffraction (right) data collection. Respectively, beamline BM29 at the ESRF synchrotron facility (Grenoble, France) and XALOC beamline at the ALBA synchrotron facility (Barcelona, Spain).

3.2.2.2.2 Nuclear Magnetic Resonance (NMR)

Nuclear magnetic resonance spectroscopy (NMR) is used to study the structure of molecules, the interaction of various molecules, the kinetics or dynamic events of molecules and the composition of mixtures of biological or synthetic solutions or composites. NMR's advantage is the unique

ability of a nuclear spectrometer to allow both the non-destructive and the quantitative study of molecules in solution and in solid state, as well as to enable the study of biological fluids. The size of the molecules analysed can range from a small organic molecule or metabolite, to a mid-sized peptide or a natural product, all the way up to proteins of several tens of kDa in molecular weight.[126]

The principle behind NMR is that many nuclei have spin (intrinsic quantum property) and all nuclei are electrically charged; rotating charged particles, acquire magnetic properties therefore interacting with external magnetic fields. Under the presence of a magnetic field, the nuclear spin for many nuclei produce an energy level splitting (Zeeman effect) at a characteristic frequency (energy). If an external field is applied, an energy transfer is possible between the base energy to a higher energy level. The energy transfer takes place at a wavelength that corresponds to the radio frequencies matching the Zeeman splitting. The signal that matches this transfer is measured in many ways and processed in order to yield an NMR spectrum for the nucleus concerned.[127]

Specifically, NMR deals with the interaction between the magnetic moments of atomic nuclei and magnetic fields. Only those nuclei for which spin is different from zero have a magnetic moment. Protein high-resolution NMR spectroscopy deals predominantly with nuclei with spin 1/2, which have no quadrupole moment and hence give rise to sharp spectral lines. The nuclei of greatest interest for organic chemistry and biochemistry – H, N, C, and P – all have isotopes with spin 1/2, though for N and C the natural abundance of these isotopes is very small, thus making necessary to label the samples with isotopes ^{13}C and ^{15}N. The proton, the most abundant nucleus in organic compounds, has (with the exception of tritium) the largest magnetic moment and hence is the simplest nucleus for investigation. NMR Spectroscopy yields spectra that provide information on the number and type of chemical entities in a molecule.[128][129]

Although each distinct nucleus experiences a distinct chemical environment and thus has a distinct chemical shift by which it can be recognized, in large molecules such as proteins the number of resonances can typically be several thousand and a 1D spectrum has overlaps. Therefore, Protein NMR utilizes multidimensional experiments to alleviate the overlapping problem. Multidimensional experiments correlate the frequencies of distinct nuclei, decreasing the chance of signal overlapping and providing more information since they correlate signals from nuclei within a specific part of the molecule. Magnetization is transferred into the sample using pulses of electromagnetic energy and between nuclei using delays. These pulse sequences allow selecting specific types of connections between nuclei. The array of NMR experiments used on proteins fall in two main categories — one where magnetization is transferred through the chemical bonds (J-coupling), and one where the transfer is through space (dipole-dipole coupling). The first category is used to assign the different chemical shifts to a specific nucleus, and the second is primarily used to generate the distance restraints used in the structure calculation, and in the assignment with unlabelled proteins.[128]

3.2.2.2.2.1 Heteronuclear Single Quantum Correlation (HSQC)

Heteronuclear Single Quantum Correlation (HSQC) is the most standard and widespread used 2D protein NMR experiment. HSQC spectra shows one peak for each H bound to a heteronucleus. In

the $^{1}H,^{15}N$-HSQC magnetization is transferred from hydrogen to attached ^{15}N nuclei via J-coupling. The chemical shift is evolved on the nitrogen and the magnetization is then transferred back to the hydrogen for detection. It shows all H-N correlations so one signal is expected for each amino acid residue with the exception of proline which has no amide-hydrogen due to the cyclic nature of its backbone. Tryptophan and certain other residues (e.g. Asn, Gln, Arg, Lys) with N-containing sidechains also give rise to additional signals. The $^{1}H,^{15}N$-HSQC is the fingerprint of a protein because each protein has a unique pattern of signal positions and is usually performed as the first heteronuclear experiment. Its analysis allows to evaluate whether the expected number of peaks is present and thus to identify possible problems due to multiple conformations or sample heterogeneity. It also assesses whether other experiment are likely to work and for instance, whether further labelling is worth investing time and economic resources.[128][130]

In the presented work:

$^{1}H,^{15}N$-HSQC experiments were recorded at 37 °C on an 800 MHz Bruker Advance III spectrometer from ^{15}N isotopically enriched protein samples in PBGD purification buffer A. 1024x200 points where collected and water suppression was achieved using WATERGATE. The data were processed using NMRPipe[131] software and in-house built scripts. Processed data were analysed with NMRDraw.

RESULTS & DISCUSSION

4

CONNEXINS
protein-protein interactions

4.1

Intercellular communication in animal tissues is almost always mediated by gap junctions (GJs). [18] GJs are the only known systems to produce direct transfer of small molecules by forming plaques of hydrophobic channels that bridge the membranes of adjacent cells. A large fraction of the communication is based on the delicate recognition mechanism between connexins and the ability of their hemichannels to ultimately generate productive intercellular channels. GJs assembly and functioning are complex, highly regulated processes and, to date, not well understood.[21]
Here, *in silico* structural modelling I have combined with *in cellulo* functional experiments to characterize connexin interactions and docking properties of hemichannels. I have also studied the role played by particular structural motives of a connexin chain and consequences brought by its mutations.

4.1.1 Intercellular gap junction channels produced by Cx26, Cx32, Cx47; a novel method mimicking docking process

4.1.1.1 Homology modelling of homomeric connexin hemichannels.

First, models for connexin hemichannel 32 (Cx32) and connexin hemichannel 47 (Cx47) were generated based on the single available high-resolution crystal structure of connexin channel 26 (Cx26) [24], which likely represents an open (i.e. active) state of the GJ [132]. A superimposition of the models is shown in Figure 4-1A and the detailed modelling protocol can be found in the Materials and Methods section (3.1.1.1).

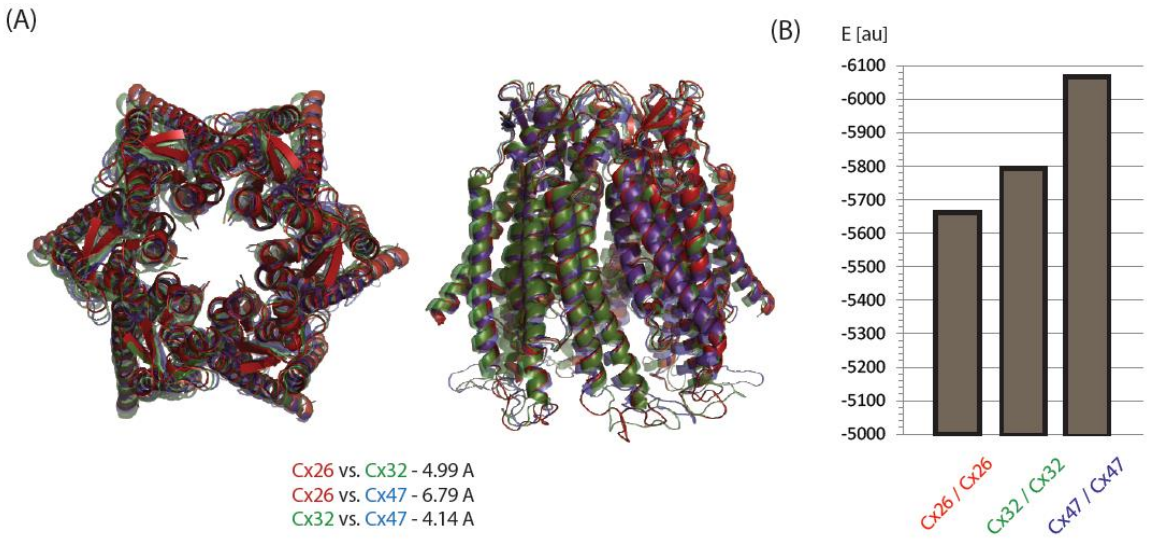


Figure 4-1. (A) Connexin 26 hemichannel (red), Connexin 32 hemichannel (green) and Connexin 47 hemichannel (blue) models with RMSD values for their superimposition. (B) Homotypic channel docking; computational results, (left) Cx26-Cx26, (centre) Cx32-Cx32, (right) Cx47-Cx47.

Only residues corresponding to regions with available structural data were modelled (corresponding to 4-218 in Cx26), while the C-terminal region of the protein (218-283 for Cx32 and 293-439 for Cx47) was not considered for modelling (Figure 3-1A). The X-ray structure of Cx26 did not show electron density for the intracellular loop region (residues 95-131) and it was also

modelled. For Cx47, the large loop adopted a conformation that resulted in steric clashes with other chains during the hemichannel assembly and it was replaced by the loop of Cx32, which shows better arrangement and higher structural homology. It is important to emphasize that both, the C-terminal and the intracellular loop regions face opposite to the interface between hemichannels, and it is unlikely for them to have an important role in the hemichannel docking mechanism. With similarities between Cx26 and the higher molecular weight Cxs of 63% (for Cx32) and 38% (for Cx47), also high level of conservations in the functionally key regions, it was reasonable to speculate that Cx32 and Cx47 may take a similar structure as Cx26.

Six Cx chains were superimposed using the bottom hemichannel of the existing structure of Cx26, to form the homomeric hemichannel models of Cx32 and Cx47. Hemichannels were subsequently combined and docked using the ClusPro server to provide *in silico* models of each intercellular GJ. The docking was carried out taking into account the hydrophobic regions of the channel and using a set of defined active residues to drive/guide the GJ formation (see Materials and Methods section, 3.1.4.1 for details). Using this procedure, the homotypic GJ models for Cx32 and Cx47 were generated and they are ranked in Figure 4-1B as a function of the specific energy requested for docking. A lower/higher energy is equivalent to a stronger/weaker complex interaction. For comparison and reference, a model of Cx26-GJ was also generated and its energy evaluated. According to the energy analysis (Figure 4-1B), homomeric Cx26-GJ would be the most stable, in agreement with experimental patch-clamp experiments that show higher intracellular Cx26 GJ yields as compared to Cx32 GJ (data not shown).

A first inspection of the GJ models reveals that the three channels show a high overall structural similarity, as expected. Figure 4-2A-C shows the hemichannel contacts distribution for the three modelled GJs. Despite the helical channel body is highly preserved among the different GJs, the interfaces of Cx26 and Cx32 and their respective interhemichannel contacts created by these GJ's are much more similar as compared to Cx47-GJ (Figure 4-2A-C).

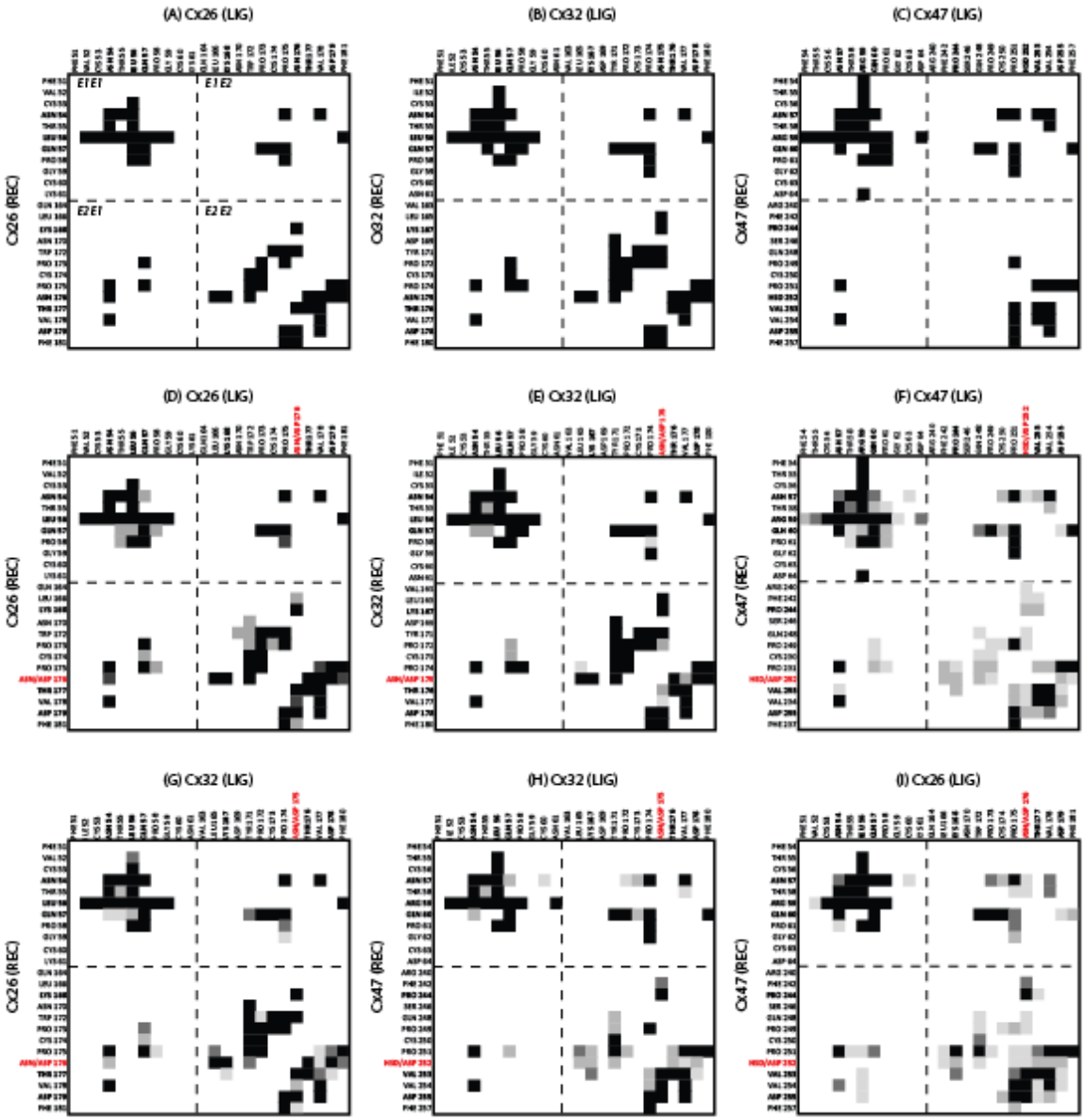


Figure 4-2. Contact maps and its changes caused by the mutation; receptor residues (listed left) vs. ligand residues (listed above the maps). Homotypic channels: (A) Cx26/Cx26, (B) Cx32/Cx32, (C) Cx47/Cx47. Changes in the contacts caused by mutations (marked red) and hemichannel composition: (D) Cx26/Cx26 channels (with & without the mutation), (E) Cx32/Cx32 channels, (F) Cx47/Cx47 channels, (G) Cx26/Cx32 channels, (H) Cx47/Cx32 channels, (I) Cx47/Cx26 channels. Black colour - contacts lees affected by the mutation, pale grey - contacts most affected by the mutation.

Contact map cannot distinguish between single and multiple contacts in the symmetric positions therefore a deeper dessimmetrized analysis of the bonding network was also analyzed. The pattern of the hydrogen bond interactions and electrostatic profiles are shown in Figure 4-3 (I, II) for Cx26 and Cx47 GJs. Consistent with the high resolution structure of Cx26-GJ[24], gap junctions are generated by the multiple intermolecular interactions generated at both, the inner (extracellular domain E1) and the outer (extracellular domain E2) rings. Nevertheless, the modelled GJs structures reveal a much diverse and rich interaction network, which is Cx dependent, in contrast to previous suggestions [111,133].

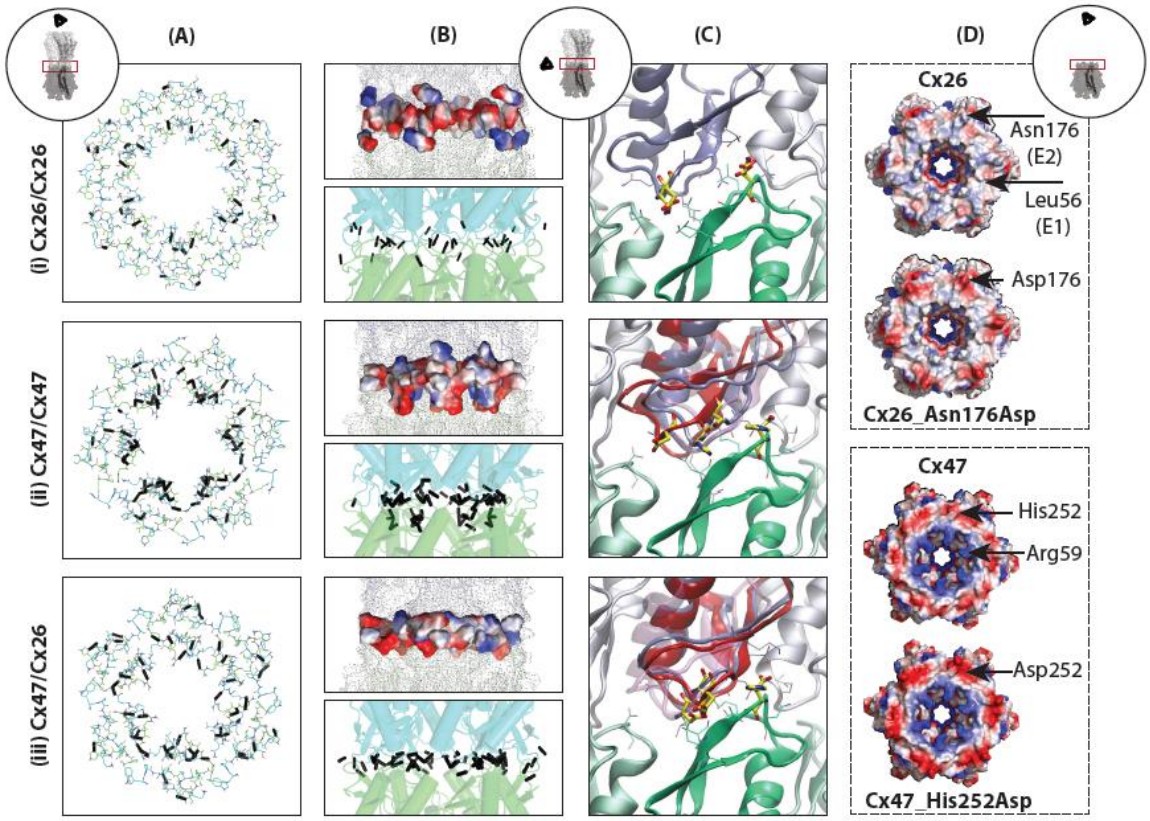


Figure 4-3. Characterization of homotypic and heterotypic channels composed by Cx26 and Cx47. (A) Hydrogen bond interactions between docked hemichannels. (B) Interactions between docked hemichannels - side view; electrostatic profile (top), hydrogen bonds (bottom). (C) Structural changes as the result of the mutation introduction; Cx26/Cx26 vs. Cx26/Cx26_Asn176Asp vs. Cx26_Asn176Asp /Cx26_Asn176Asp (I), Cx47/Cx47 vs. Cx47/Cx47_His252Asp vs. Cx47_His252Asp /Cx47_His252Asp (II), Cx47/Cx26 vs. Cx47/Cx26_Asn176Asp vs. Cx47_His252Asp /Cx26_Asn176Asp (III). (D) Electrostatic profile of the hemichannel docking interfaces: Cx26_WT and Cx26_Asn176Asp (top), Cx47_WT and Cx47_His252Asp (bottom).

Figure 4-3D shows the electrostatic surfaces for Cx26 and Cx47 hemichannels within the functional GJ model. GJ's obtained from Cx26 (and – similarly – Cx32) do not suggest a major role from electrostatic interactions since their interfaces are rather neutral. Instead, Cx47-GJ shows charge polarization between the inner (E1) and the outer (E2) rings caused by the existence of long, positively charged Arg59 instead of the hydrophobic Leu56 that occupies the corresponding place in Cx26 and Cx32. We hypothesize that these differences lower the stability for the Cx47-GJ and may help explaining the lower efficiency of Cx47 in forming GJs as compared to the Cx32- and Cx26-GJs.

We have also analysed the hydrogen bond network generated between the two hemichannels for the different GJ models (Figure 4-3 I-II, A-B). GJs generated from Cx26 (and – similarly – Cx32) show a moderate density of intermolecular H-bond interactions, almost evenly distributed across the interface with a stronger contribution for the inner ring E1. For Cx47-GJ, this tendency is clearly accentuated and E1 concentrates almost all the intermolecular H-bond interactions, which largely exceed the number found in the other GJ models. Such density of intermolecular interactions may help overcoming the electrostatic repulsion found in Cx47-GJ.

4.1.1.2 Extension of the structural model to GJ channels involving heterotypic and mutant connexins.

Connexins have the ability of recognizing different chains, thus enabling the generation of heterotypic GJs that are conformed by homomeric (e.g. Cx26/Cx26_MUT, Cx26/Cx32) and heteromeric (e.g. in co-injection experiments) hemichannels. Following the same strategy as previously described, we have computationally generated structural models for all possible heterotypic hemichannels and docked all the combinations involving the three Cxs under consideration (Figure 4-4A).

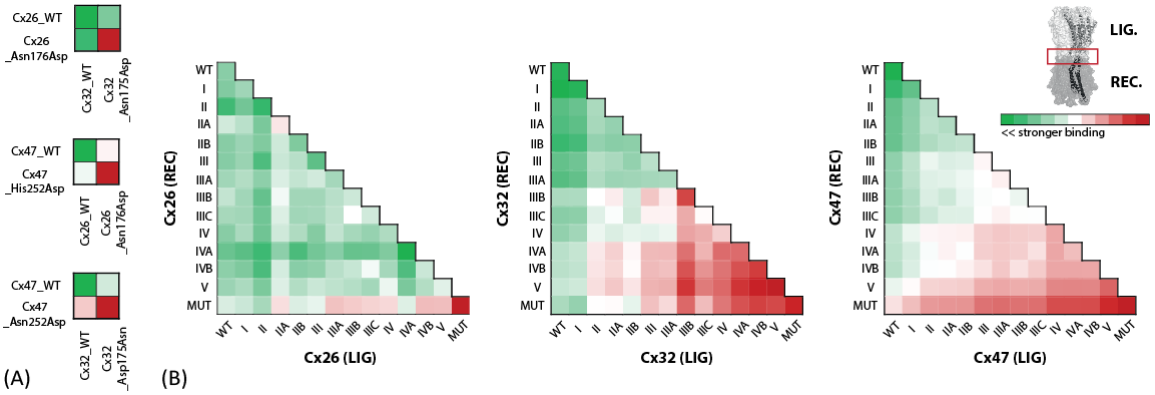


Figure 4-4. (A) Docking energy for homotypic and heterotypic channel types composed of Cx26 and Cx32 (top), Cx47 and Cx26 (centre), Cx47 and Cx32 (bottom) with their mutations. (B) Docking energy for all the channel types composed of Cx26 (right), Cx32 (centre), and Cx47 (left) with their mutations Cx26_Asn176Asp, Cx32_Asn175Asp and Cx47_His252Asp, respectively.

According to the resulting docking energies, Cx26/Cx32-GJ has a significantly stronger interaction energy (-5800) than Cx26/Cx47- (-5600) or Cx32/Cx47- (-5620) GJs and, consequently, the model predicts a superior GJs in terms of docking energy as compared to the other two heterotypic arrangements. Consistently, the existence of heterotypic Cx26/Cx32-GJs has been experimentally observed [112,134], while no experimental evidence can support the existence of the other two heterotypic GJs.

Analysis of the GJ models underscore the importance of intermolecular H-bond interactions for the stability of the GJ complex (*vide infra*). In Cx26, the crystal structure available allowed identifying Asn176 as a key residue for the connexon of hemichannels [24], in synthony with the pathogenic effect observed in the Asn176Asp mutation (or corresponding in Cx32 and Cx47) [111]. The GJ models generated in this study indicate that an equivalent role may be played by Asn175 in Cx32, and by His252 in Cx47-GJs.

Here, we have investigated the energetic role of these pivotal residues by modelling the mutant connexins and GJs in an equivalent manner as for the WT GJ models. To that end, all the possible combinations between WT and mutant (Asn176Asp, Asn175Asp and His252Asp for Cx26, Cx32 and Cx47 respectively) chains have been generated (Figure 3-2) and their stabilities (obtained from the in silico docking energy) have been compared (Figure 4-4B). It is important to emphasize that most of the complexes studied have no ultimate physiological implantation, but interesting conclusions emerge from the collective analysis of the relative energies for the different GJs under

consideration. In there, the energies have been normalized (see Materials and Methods, 3.1.4.1) and they are ranked according to their stronger (green) or weaker (red) capacity to form a GJ.
Figure 4-4B shows the relative energies for the different heterotypic channels with a progressive incorporation of mutant connexin chains. Cx26-GJ are rather tolerant to the incorporation of the mutant side chains and GJ formation is disfavoured only when a large number of mutations are interacting (Figure 4-4B left). This is attributed to specific arrangements of mutations (i. e. the alternation of mutations in one hemichannel) that particularly penalize the GJ formation in a cooperative mechanism. On the other hand the energy patterns observed for Cx32 (WT/Asn175Asp)-GJs (Figure 4-4B centre) and Cx47 (WT/His252Asp)-GJ (Figure 4-4B right) agree well with a progressive energy reduction upon mutation accumulation, but with no significant energy discrimination among the different topological arrangements. This result is consistent with the electrostatic nature observed for Cx47-GJ: mutations progressively increase the electrostatic repulsion at the interface, regardless of their specific location.
Finally, we also have noticed that an asymmetric mutated chain rearrangement is energetically favourable: the higher the asymmetry in the mutation location, the bigger the chance to create an optimal interaction network for a functional channel formation.

4.1.1.2.1 Structural analysis of the channels produced by Cx26, Cx32, Cx47

All the modelled channels exhibited a marked division between two circles of interactions: the inner circle (created by residues of extracellular domain E1) and outer (extracellular domain E2). Moreover, the structural models confirms the previously reported responsible interactions to form GJs with Cx26 and Cx32 [24,111]: the crucial residues forming the net of interaction in channels are Asn54, Leu56, Gln57 (domain E1) and Lys168, Asn176, Thr177, Asp179 (domain E2), Lys167, Asn175, Thr176, Asp178 in Cx32. However, we do not observe interactions between hemichannels for Gln57 nor between Asn176 and Asp179, while the interaction between Tyr171 and Asn/Asp176(175) is widely represented whenever a Cx32 chain appears in the system. The similar interaction patterns are observed between Cx26/Cx26- and Cx32/Cx32-GJ (Figure 4-2A-B), consistent with the high yields obtained in cellular experiments involving these connexins.
In contrast, GJs obtained from Cx47 hemichannels reveal a very different contact map (Figure 4-2C). Actually the total number of interfacial interactions decreases, again in line with the lower yields obtained in patch-clamp experiments. Most of the interfacial interactions are now concentrated in the domain E1, with positively charged (Figure 4-3D), long Arg59 (this residue corresponds to Leu56 in Cx26 and Cx32: Leu56$^{Cx26,\ 32}$) being responsible for most of the interactions. This residue gets deep into coupled connexons and is also responsible for smaller pore diameter of Cx47 hemichannels (Figure 4-5B). Other contributing residues are: Cys56 (53^{Cx26,32}), Asn57 (54^{Cx26,32}), Thr58 (55^{Cx26,32}), Gln60 (57^{Cx26,32}), Asp64 (Lys61^{Cx26} and Asn61^{Cx32}). On the other hand, E2 ring provides little interactions to Cx47/Cx47-GJ (Figure 4-2C). His252 in Cx47, corresponding to Asn176^{Cx26} & Asn175^{Cx32}, does not create many hydrogen bonds, while Asp252 in Cx32 and Cx26 does.

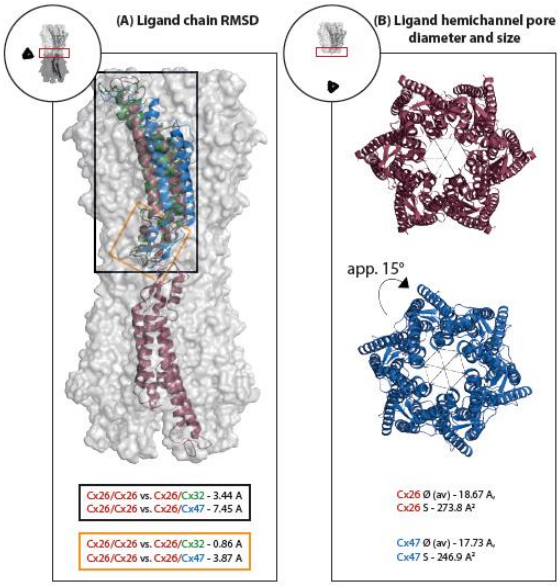


Figure 4-5. Characterization of homotypic and heterotypic channels composed by Cx26 (red) and Cx47 (blue). (A) Ligand chain RMSD; additionally marked a connexin compatible to Cx26, Cx32 (green). (B) Ligand hemichannel pore diameter and size.

In contrast to the mutation Asn176^{Cx26}/175^{Cx32}Asp that destroys interaction net of the residue 176(175) but it does not affect much the E1 interaction net (Figure 4-2D-E), the corresponding mutation His252Asp in Cx47 creates new interactions with multiple residues, disturbing residues in E1 (hydrogen bond amount drops significantly) which are the core of the channel stability (Figure 4-2F). The detailed analysis of the hydrogen bond network revealed the appearance of interactions between Asp252 (176^{Cx26}, 175^{Cx32}) and Arg240 (Gln164^{Cx26}, Val163^{Cx32}) and between Gln248 (Trp172^{Cx26}, Tyt171^{Cx32}) or Val253 (Thr177^{Cx26}, 178^{Cx32}) with Asp255 (179^{Cx26}, 178^{Cx32}). On the other hand, other interactions present in wild type GJs are significantly reduced upon mutation: Val253-Val253, Asn57 with Thr58, Arg59, Gln60, Pro251 (175^{Cx26}, 174^{Cx32}) as well as Arg59 with Cys56, Thr58 and Asp64. Thus, when the mutation disrupts the hydrogen bonds created by residue 176 (or corresponding) other hydrogen bond appeared to rescue the system (Figure 4-6, little green circles) ultimately altering the hydrogen bonding network.

Residue couple			Hydrogen bond (occupancy per channel)		
REC.		LIG.			
ASN54	OD1	LEU56-N	100%	100%	67%
	ND2	ASN54-O	67%	83%	100%
	O	ASN54-ND2	33%	50%	67%
THR55	OG1	ASN54-OD1	2%	17%	0%
LEU56	N	ASN54-OD1	100%	100%	100%
	O	GLN57-N	0%	0%	17%
LYS168	NZ	ASN/P176-OD1/2	67%	50%	33%
ASN/P176	OD1	LYS168-NZ	100%	67%	50%
	ND2	ASP179-OD1	17%	0%	0%
ASP179	OD2	ASN/P176-ND2	0%	33%	0%

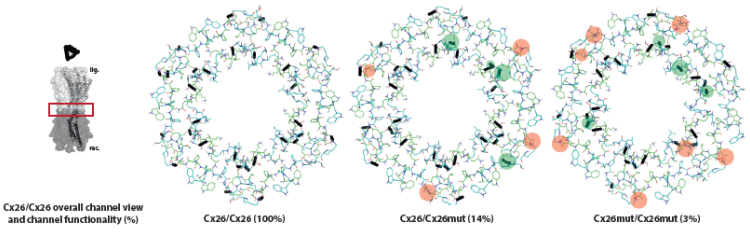


Figure 4-6. Hydrogen bonds for Cx26/Cx26 channel (left), Cx26_Asn176Asp/Cx26 channel (centre), Cx26_Asn176Asp /Cx26_Asn176Asp channel (left). Table - hydrogen bond occupancy (%, computational results), green colour – higher, white – lower. Cannel ross-sections – hydrogen bond net, little red circles – interactions affected (disrupted) by the Asn176Asp mutation, green circles – new interactions appearing after the mutation, channel functionality (%, experimental results).

We hypothesize that a channel doesn't need exactly 6x10 interactions as it was previously proposed [111], but a lower number of interactions between hemichannels, provided they are distributed across the interface, is sufficient to produce a stable GJ. Nevertheless, if such compensation is not possible, the GJ will not be formed, thus providing a structural explanation for the deleterious effect of these mutations observed in vivo. In like manner, homomeric heterotypic channels created of Cx47 and Cx26 (or – similarly – Cx32) exhibit change distribution of atomic contacts (compare: Figure 4-2A-C vs. Figure 4-2H-I), visible in more details in Figure 4-3III (vs. Figure 4-3I-II).
An interesting observation is the asymmetry which appeared in all the docked complexes. According to the models, the interaction net is not as rigid nor symmetric as it was proposed before, with up to 60 hydrogen bonds per channel [133]. This effect seems to be produced because there is not exactly the same number of hydrogen bonds between paired chains in a channel (Figure 4-3, Figure 4-6). Asymmetry in the GJs assembly has already been suggested in the literature [24]: Maeda *et al.* claim that E1 in one hemichannel interacts with two residues belonging to the E1 rings from the other hemichannels, whereas one residue from the E2 domain interacts with one residue from the E2 ring of the opposite hemichannel, ultimately resulting in a 30° tilt between hemichannels. In our models, however, one chain from the top hemichannel couples with one from the bottom hemichannel (Figure 4-5). A detail that distinguishes Cx47 hemichannel, is a slight turn of about 15° in the relation to the opposite hemichannel (Figure 4-5) when docked to Cx26(Cx32).

4.1.1.3 Model validation using experimental data.

The computational models were submitted to experimental validation at different levels and the summary is shown in the Figure 4-7:

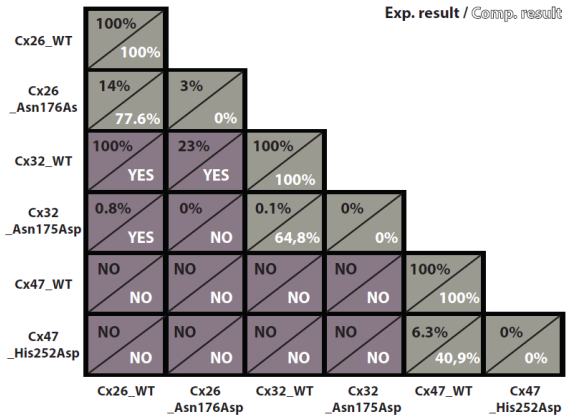


Figure 4-7. The experimental and computational results comparison. Grey colour – channels made of the same connexin type, violet – channels made of different connexin types.

First, we found that the relative docking energy correlates well with the experimental yield for the formation of the Cx26-, Cx32- and Cx47-GJs (Figure 4-1B). This validation can be extended to the heterotypic GJs, and we found that the docking energy is a bona fide reporter of the ability of connexins to create GJs. Next, we also calculated the relative energies for the combination of the mutant variants and WT connexins (Figure 4-4). We observed that the incorporation of the mutation reduced the docking energy, resulting in less stable GJs. The relative energies for the

different GJ combinations, represented by the colour code in Figure 4-4A, agree well in general with the experimental results: for the three Cx, no experimental evidence for the Cx(Mut/Mut)-GJ is reported, in full agreement with the low docking energy values. For the heterotypic connexins, the experimental yields of GJ formation (relative to WT) are 14%, 0.1% and 6.3% for Cx26(WT/Asn176Asp)-, Cx32(WT/Asn175Asp)- and Cx47(WT/His252Asp)-GJs respectively. The calculated values for the docking energy are 77%, 64% and 41% providing moderate agreement with the experimental data. Hence, since the docking energy reflects the strength of the interactions between hemichannels, the good correlation found adds confidence to the structural models reported here.

In a second and more stringent experiment, patch clamp experiments co-injecting Cx47(WT) and the mutant under consideration Cx47(His252Asp) were performed in Xenopus oocytes (see Materials & Methods for details). Co-injection allows the free assembly of the connexins and, for a junctional channel, the number of theoretical combinations is 196 (14^2), assuming there are no rotary or structural preferences [23]. Then, for one type connexin with its mutant – omitting the analogue (reflected) systems – there are 105 possible forms. The schematic representation of the 14 different hemichannel forms with the probability of its occurrence is presented in the Figure 3-2. Thus, co-injected connexins will generate a "mix" of GJs and the experimental yield of such junctions in the patch clamp experiment are reported relative to Cx47(WT/WT)-GJ (referred to 100%) and Cx47(His252/His252Asp)-GJ (that does not produce any junction and it corresponds to the 0% threshold). The yields oscillate between 50% for the combination Cx47(WT/mix)-GJ to less than 10% form the Cx47(WT/His252Asp)-GJ (Figure 4-8).

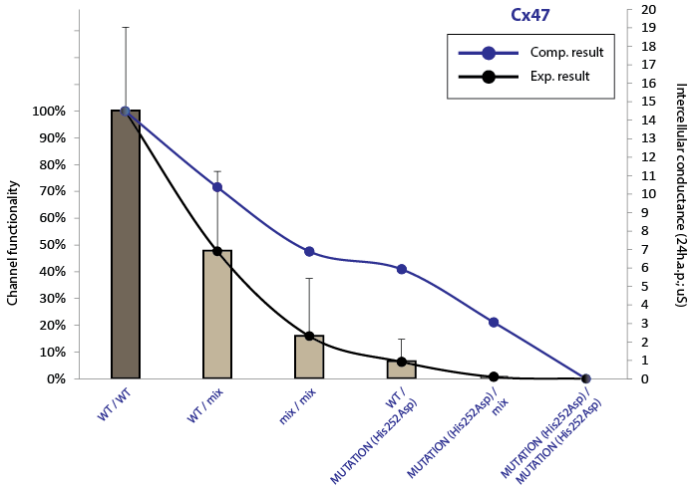


Figure 4-8. Homomeric and heteromeric channel docking; experimental and computational results comparison for Cx47. Dark grey – WT/WT (homomeric homotypic channel).

To predict the docking energies, we have generated computational models for all the possible GJs (144) and averaged the individual energies according to the combinatorial rules described above and normalized with respect to the Cx47(WT/WT)-GJ docking energy. As shown in Figure 4-8 (blue dots), the calculated energies agree well with the experimental yields, being able to correctly predict the relative capacity of a given combination to form a GJ, thus validating the structural models employed for the calculation.

4.1.2 Gap junction channels produced by Cx26; mutations in E1 and E2 extracellular domains help to gain insight in the docking mechanism and its disturbances.

Following the established protocol of modelling and docking procedures, four additional channels were generated: Asn54His, Asn54Ile, Asn54Lys, Asn54Ser. As it is believed, both extracellular loops are associated with cell–cell recognition and docking [20], although it was suggested that they play different roles. Protein sequence alignment analyses on the docking compatible and incompatible connexins indicated that the E1 domain may be important for the formation of the GJ channel (stability) and the E2 for the docking compatibility (selectivity). Furthermore, the hydrogen-bond forming residues in both E1 and E2 domains are mutational hot spots for connexin linked human diseases [66]. For instance, four hearing-loss associated Cx26 mutations were found on Asn54: Asn54His, Asn54Ile, Asn54Lys, Asn54Ser [135–139] and these mutations have been studied in the present thesis.

4.1.2.1 Asparagine 54 of connexin 26 is a key residue for the stability of the homomeric homotypic gap junction channel docking.

Preliminary results of functional assays show that Cx26_Asn54Ser, but not Cx26_Asn54Ile nor Cx26_Asn54Lys (fewer experimental evidence in this case), retains the ability to form functional hemichannels and gap junction channels. Interestingly, isolated oocytes expressing Cx26_Asn54Ser induce higher levels of hemichannel currents (2-3-fold more than Cx26_WT) and the Cx26_Asn54Ser-Cx26_Asn54Ser pairs huge levels of electrical coupling (up to 5-10-fold higher than those detected between the Cx26_WT- Cx26_WT pairs); data yet unpublished. To examine the structure-function relationship at the sub molecular domains and individual amino acid residues level, we have generated computational models for the GJs channels and compared the docking energies obtained for all the complexes. Results are shown in the Figure 4-9:

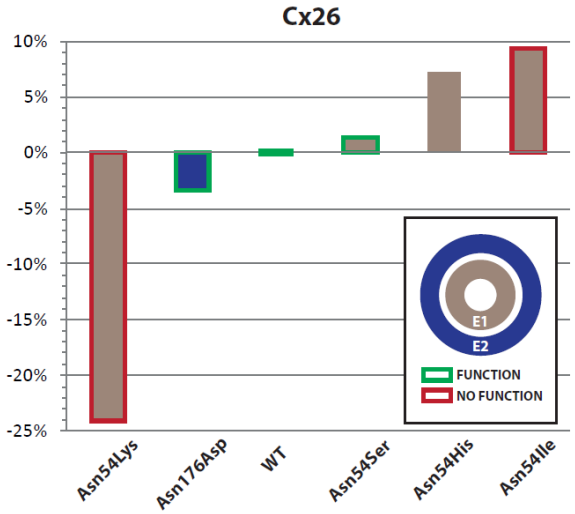


Figure 4-9. Homotypic channel docking; difference in the docking energy (%) between (from left to right): Cx26_Asn54Lys-Cx26_Asn54Lys, Cx26_Asn176Asp-Cx26_Asn176Asp, Cx26_WT-Cx26_WT (reference complex), Cx26_Asn54Ser-Cx26_Asn54Ser, Cx26_Asn54His-Cx26_Asn54His, Cx26_Asn54Ile-Cx26_Asn54Ile. Brown colour – E1-mutant, blue – E2-mutant, green mark/square – complex experimentally reported as functional, red square – not functional, no square – no information.

Cx26_WT-Cx26_WT is represented as a green line and same, the reference. We attribute a lower/higher docking energy to a weaker/tighter association, using the WT as a reference. According to the experimental results, only Cx26_Asn176Asp-Cx26_Asn176Asp and Cx26_Asn54Ser-Cx26_Asn54Ser form functional channels with 3% and much less (data yet not published) of the WT conductivity, respectively (both marked with green square in the Figure 4-9), while Cx26_Asn54Lys-Cx26_Asn54Lys and Cx26_Asn54Ile-Cx26_Asn54Ile do not. The most straightforward interpretation is that a channel requires a minimum energy to form a stable system and mutated hemichannels can form a functioning channel only if their energy does not differ much from the one of WT channel. Following this logic, the Cx26_Asn54His hemichannels – for which no experimental data is available – are predicted not to form a functional channel. The fundamental open question is how such energy value can be correlated with the interface interactions between hemichannels. To investigate this point, additional analysis were made. Figure 4-10 shows atomic contacts between hemichannels for the different GJs under consideration. In each case changes are notorious, proving that even a point mutation spreads over the entire docking interface. This is not surprising, since a minor change in a connexin chain will be amplified 6 times in the hemichannel and 12 times in the whole gap junction channel.

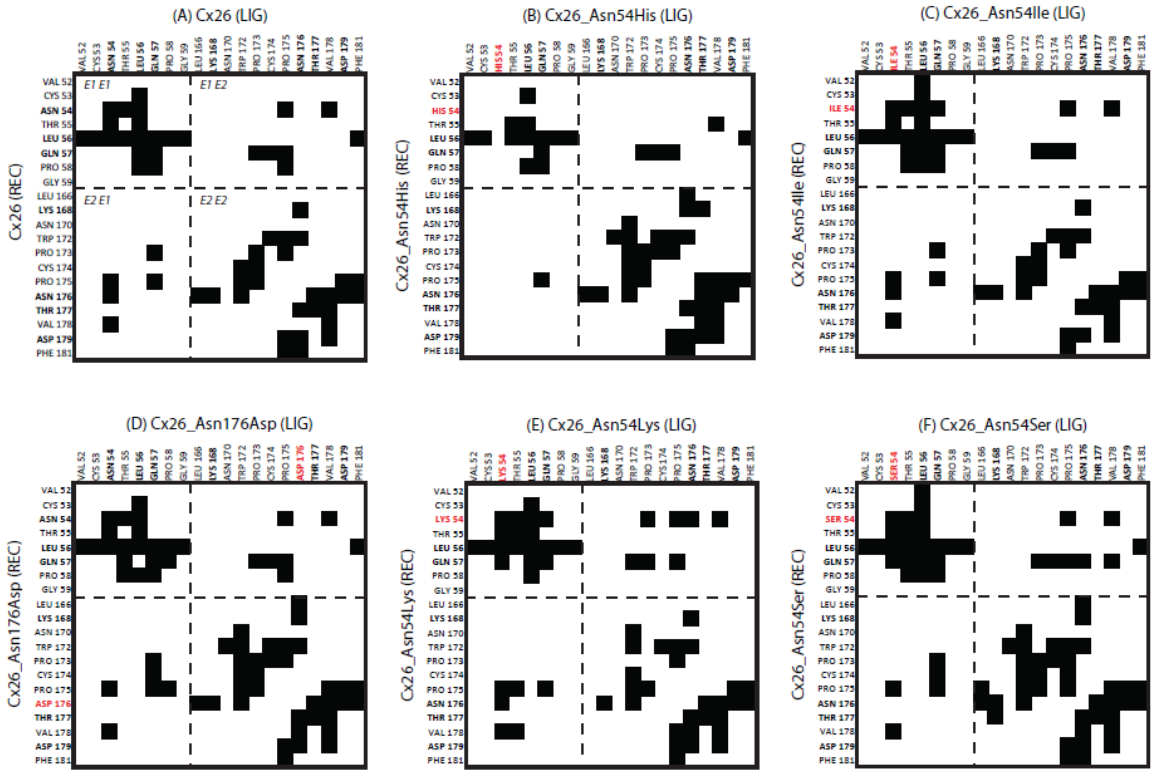


Figure 4-10. Contact maps; receptor residues (listed left) vs. ligand residues (listed above the maps). Homotypic channels: (A) Cx26_WT-Cx26_WT, (B) Cx26_Asn54His-Cx26_Asn54His, (C) Cx26_Asn54Ile-Cx26_Asn54Ile, (D) Cx26_Asn176Asp-Cx26_Asn176Asp, (E) Cx26_Asn54Lys-Cx26_Asn54Lys, (F) Cx26_Asn54Ser-Cx26_Asn54Ser. Mutations (marked red).

Atomic contacts cannot be directly attributed to hydrogen bonds. These were - respectively – 30 for the WT channel which contacts are represented in (Figure 4-10A), 22 for (B), 6 (C), 26 (D), 26 (E), 24 (F). The detailed analysis of key Cx26_WT-Cx26_WT interactions and the effect of Asn176Asp mutation can be found in the 4.1.1.2.1 subsection. In case of Asn54His mutation (B) (in comparison

to WT channel) there were no mutual interactions between His54 residues. In the case of Asn54Ser (F), most of hydrogen bonds were concentrated in the E1-E1 area, where Ser54, Thr55, Leu56 and Gln57 and they were found interacting with each other. The opposite situation was reported in case of Asn54Lys (E) where most hydrogen bonds were concentrated between E2 domains, with Lys168, Asn176, Thr177, Val178 and Asp179 found forming interactions. No mutual interactions between residues 54 could be spotted, similar to the situation described in (B). The Asn54Ile-mutated channel (C) showed hydrogen bonds only between Asn176 and Lys168.

Although previous observations had already confirmed the crucial role of residue Asn176 for the proper hemichannel docking, the results presented here add valuable information and support the hypothesis that the E1 domain – with the pivotal role of Asn54 – is ultimately responsible for the GJ channel stability.

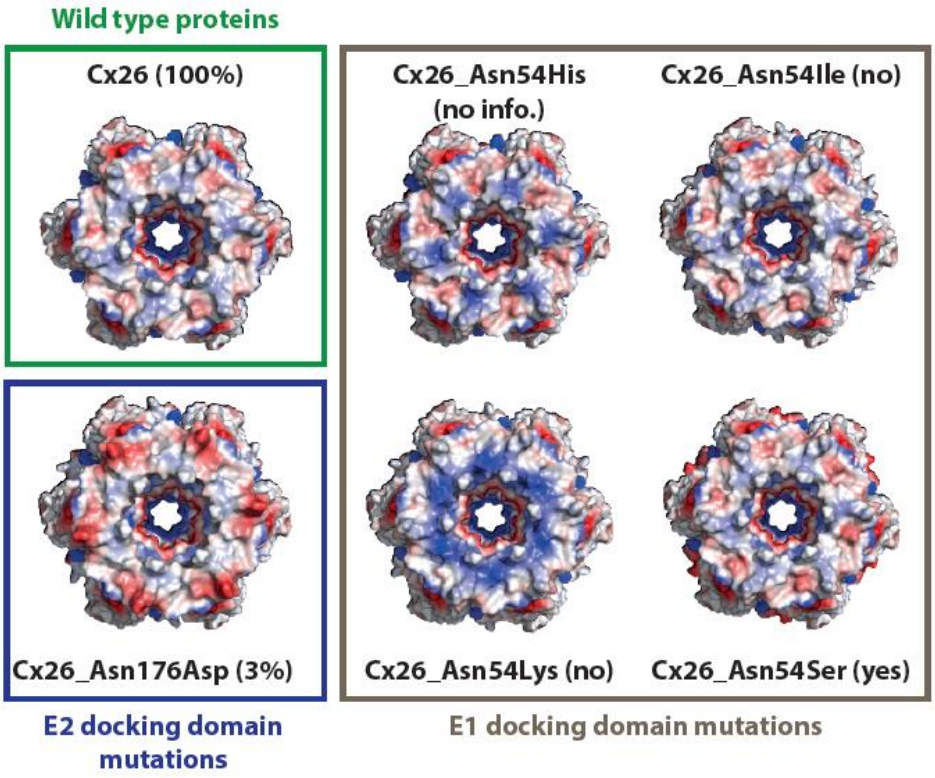


Figure 4-11. Electrostatic profile of the hemichannel docking interfaces; comparison between Cx26_WT hemichannels and the mutants.

Focusing now on the on the interface electrostatics analysis , it is clear that Asn54His/Lys results in too many positive charges at the interface, which may repel the neighbour hemichannel (Figure 4-11). In other words, the electrostatic repulsion may significantly raise the energy to form a stable channel., Moreover, Asn54His/Lys loses the key Asn54-Asn54 stability interaction, in contrast to Asn176Asp, that is still able to form a functional channel despite the charged network. In the case of Cx26_Asn54Ile-Cx26_Asn54Ile, the experimental evidence that proves no channel formation cannot be explained by any significant charge change. However, Asn is a small, polar-uncharged amino acid, which is replaced by Ile, a non-polar, branched and very hydrophobic residue, which tends to be buried inside the protein core (Figure 4-12).

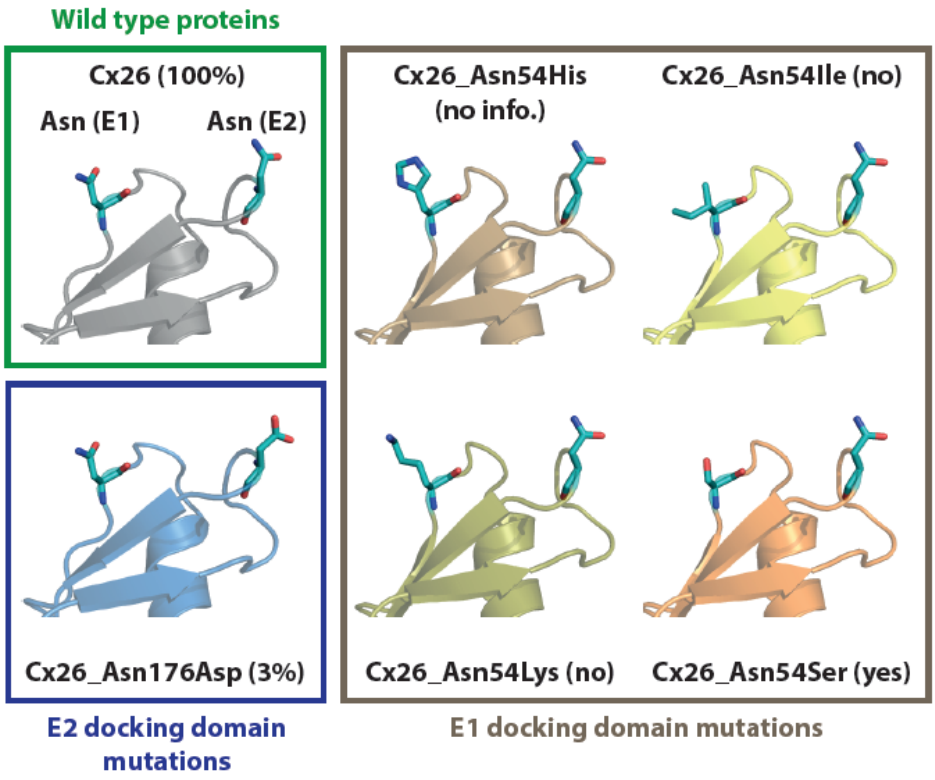


Figure 4-12. Structural differences between WT Cx26 hemichannel and the mutants.

In similar terms we explain the Cx26_Asn54Ser-Cx26_Asn54Ser functionality. Serine has a reactive hydroxyl group on its side chain that can easily interact with other residues. It also can form covalent bonds to other substituents, modifying the protein structure. Serine, as asparagine, is one of the amino acids commonly occurring at protein surfaces.
In conclusion, the docking energy value itself – taking in the account limitations of the model (e.g. lack of membranes) and method (rigid docking) – has to be considered as a relative reporter; it does not reflect any real characteristics or property of a system but it reports on the *changes* that arise in atomic compositions of a system.

4.1.3 Molecular dynamics simulations on Cx47 hemichannels prove importance of N-terminal and transmembrane domains for the proper functioning of gap junction systems.

Continuing the issue raised in the previous section about different roles played by different structural motives of a connexin chain, we have also analyzed the N-terminal (NT) and transmembrane (TM) domains. It is believed that they generate hemichannel pore, and thereby they have an impact on hemichannel functionality, with particular emphasis on gating mechanism [20]. Gating regulation is associated with a variety of factors that can differ depending on the connexin subtypes that compose the channel. Functional and electrophysiology studies have provided several mechanisms to explain this process (Figure 4-13). However, structural models supporting these findings are still not sufficient.[140]

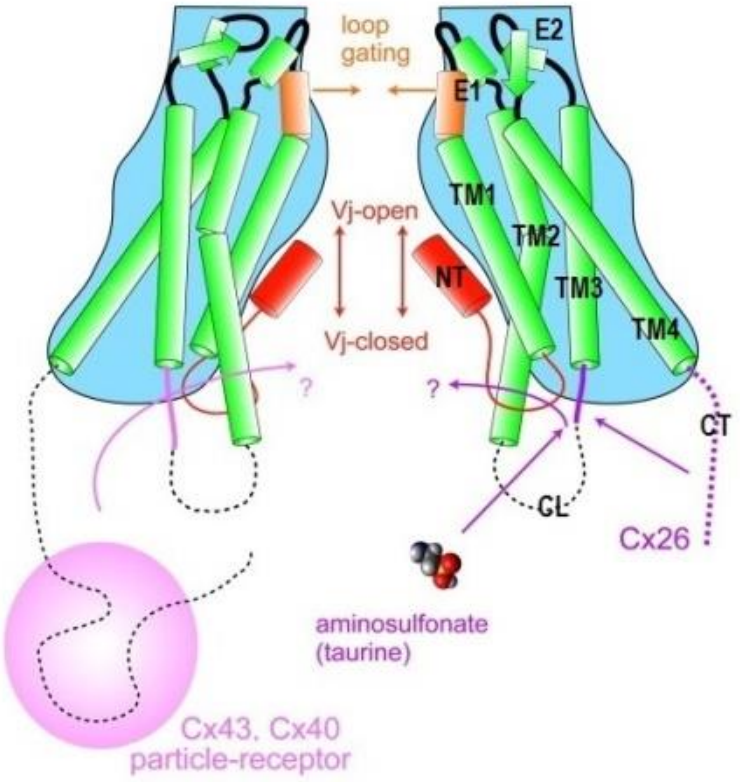


Figure 4-13. Types of hypothesized movements for gap junction gating. [140]

The most important parameter that regulates the opening and closing of intercellular channels and undocked hemichannels is, certainly, voltage. Most connexin channels display two distinct voltage-gating processes: Vj-gating (fast gating) and loop-gating (slow gating). Both properties are intrinsic to hemichannels, operating separately by the actions of different voltage sensors and gates [141]. Vj-gating type (marked red in the Figure 4-13) is initiated by the translocation of charges in the N-terminus [142]. Loop-gating type (orange) involves the permeability barrier formed by narrowing the channel pore of the TM1/E1 domain and is the mechanism that most probably keeps undocked hemichannels closed in plasma membranes [143]. Besides, there are other hypothesized types like the particle-receptor model type (pink), proposed for Cx43 and Cx40 junction channels involving an interaction between CT and CL loops (CL as a receptor) [144], and the modulation type (purple), represented here by taurine that binds to CL and involves CT loop [145].

Cx47 highly expresses in oligodendrocytes and it is used in the present work as an example to study the effect of modifications in the gating responsible regions. Mutations in the GJC2 gene, encoding Cx47, cause the autosomal recessive Pelizaeus–Merzbacher-like disease – neurological disorder of cerebral hypermyelination, which is clinically characterised by impaired motor development, ataxia, choreoathetoid movements, dysarthria and progressive limb spasticity [146,147].

Three altered hemichannel models (Cx47_NT_{TR}, Cx47_ Glu44Asn and Cx47_Tyr47His) were generated, subjected to MD simulation and compared to the Cx47_WT hemichannel. The trajectories collected during the sampling time (8 out of 10ns each) were submitted to qualitative and quantitative analysis, including a visual inspection of the trajectories and average structures and calculation of lifetime hydrogen bonds. Detailed protocols and verification of the Cx47

hemichannel modelling can be found in the 3.1.1.1 and 4.1.1.1 sections, detailed MD protocol – in the 3.1.3 section.

4.1.3.1 Cx47 modifications' influence on the pore net charges.

Cx47_NT$_{TR}$ stands for Cx47 protein with three first amino acids truncated at the NT domain. Functional data found a 2-fold increase in ionic conductivity upon NT truncation. However, long-NT channels show cationic dye selectivity, while short-NT channels do not discriminate between positive and negative dyes, even considering that the electrostatic pore profile remains unchanged. Glu44Asn mutation model was added for verification purposes due to the fact that first analysis strongly suggested an interaction of NT domain with Glu44. According to experimental studies, Cx47_Tyr47His mutant is able to form functional solitary hemichannels with similar properties as the full-length Cx47_WT. However, these mutated hemichannels are not able to dock neither between themselves nor with Cx47_WT hemichannels to form complete intercellular channels. Finally, Tyr47 of Cx47 was defined previously as the first residue of the E1 domain (Figure 3-1C).
NT domain of each connexin molecule folds into the gap junction pore and contributes to the lining of the pore inner surface. In addition, six NT domains form a funnel structure at the pore entrances which is the narrowest part of the (hemi)channel. Surface electrical charges at the site of NT domains and small physical pore diameter generate electrical and physical barriers to ions going through the (hemi)channel, respectively. This kind of arrangement provides a structural advantage for the NT domain in sensing a voltage change across the channel and/or plays a role in limiting the rate of ion permeation through a channel, a parameter that can be quantitatively measured as the single (or unitary) channel conductance [148].
The net charge number at the NT domain has been suggested in the literature to be an important factor in channel conductance and in V_j-gating [148], therefore, we have introduced labelled charge residues in its beginning portion (Figure 4-14).

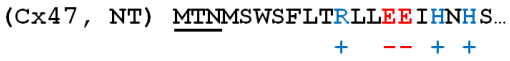


Figure 4-14. Net charges of the Cx47 N-terminal domain. Blue – positively charged residues, red – negatively charged residues, underlined text – truncated residues.

Among the first 20 amino acids, there are 5 charged residues (3 positive and 2 negative), with a net theoretical charge of +1. Considering hexameric connexin oligomerization to form one hemichannel, the total theoretical net charge of the NT domains of the Cx47 hemichannel would be +6. Taking into account that none of the charged residues locate in the area of truncation (neither close), the truncation analysis will focus on the intermolecular interactions and the subsequent structural alterations.
On the other hand, the loop-gate permeability barrier is suggested to be constituted by a segment of the first extracellular loop (E1 domain) accompanied by straightening of the bend angle between E1 and the first transmembrane domain (TM1). Dynamic fluctuations in interacting van der Waals and electrostatic network in each subunit are directly linked to the stability of this structure.

Similar than in the previous case, the electrostatic network includes charged residues that are pore-lining and thus positioned to be voltage sensors [141]. They are presented in the Figure 4-15.

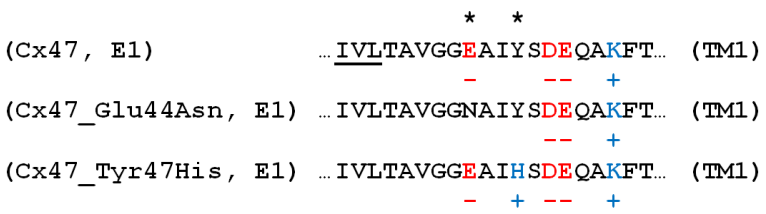


Figure 4-15. Net charges of the Cx47 chain in the E1/TM1 area . Blue – positively charged residues, red – negatively charged residues, mutation spots marked with asterisks.

In contrast to NT truncation, Glu44Asn and Tyr47His mutants bring substantial changes to the net charge, increasing it from -2 to -1 (per chain). However, it is very unlikely that they are so pivotal for the channel pore function. To elucidate the mechanism of connexin gating with atomic resolution, the altered Cx47 hemichannels were structurally and energetically compared to the WT structure. Particular attention was drawn to the dynamics of the interactions that stabilize the pore, pore diameter and electrostatics profiles of the pore.

4.1.3.2 Structure and dynamic features of the Cx47 hemichannel and its altered versions.

MD simulations allowed relaxing the Cx47 hemichannel structure (based on the Cx26 crystal structure) resulting in an expansion of the pore and some changes in the packing of the transmembrane helices. Notably, the extracellular domains almost did not change. Hydrogen bonds were monitored along the trajectories and interactions between Met1, Thr2, Asn3, Glu/Asn44 and Tyr/His47 that lasted longer than 5% of the simulation time are listed in the Table 4-1 and showed graphically in the Figure 4-16.

Residue couples			Hemichannel			
			Cx47_WT	Cx47_NT$_{TR}$	Cx47_Glu44Asn	Cx47_Tyr47His
No.	Acceptor	Donor	Hydrogen bond occupancy per chain			
1	MET1	MET1	**8,52%**	0,00%	1,79%	0,00%
2	MET1	GLU/ASN44	**37,85%**	0,00%	5,21%	51,50%
3	THR2	THR39	**0,00%**	0,00%	19,31%	6,54%
4	ASN3	ASN3	**15,33%**	0,00%	16,90%	2,29%
5	ASN3	MET4	**5,63%**	0,00%	1,79%	8,90%
6	TRP6	ASN3	**1,52%**	0,00%	10,48%	17,11%
7	GLU/ASN44	ALA40	**5,08%**	4,10%	5,48%	4,04%
8	TYR/HSD47	GLU188	**0,00%**	7,40%	3,46%	0,00%
9	TYR/HSD47	LYS189	**16,25%**	13,35%	13,79%	1,38%
10	TYR/HSD47	GLY42	**19,19%**	10,50%	19,50%	0,42%
11	SER48	GLU/ASN44	**13,02%**	3,98%	5,21%	12,35%
12	ARG78	TYR/HSD47	**0,00%**	1,48%	7,15%	7,75%
13	ARG185	TYR/HSD47	**0,54%**	18,40%	0,00%	7,27%

Table 4-1. *Hemichannel interactions affected by chain modifications.*

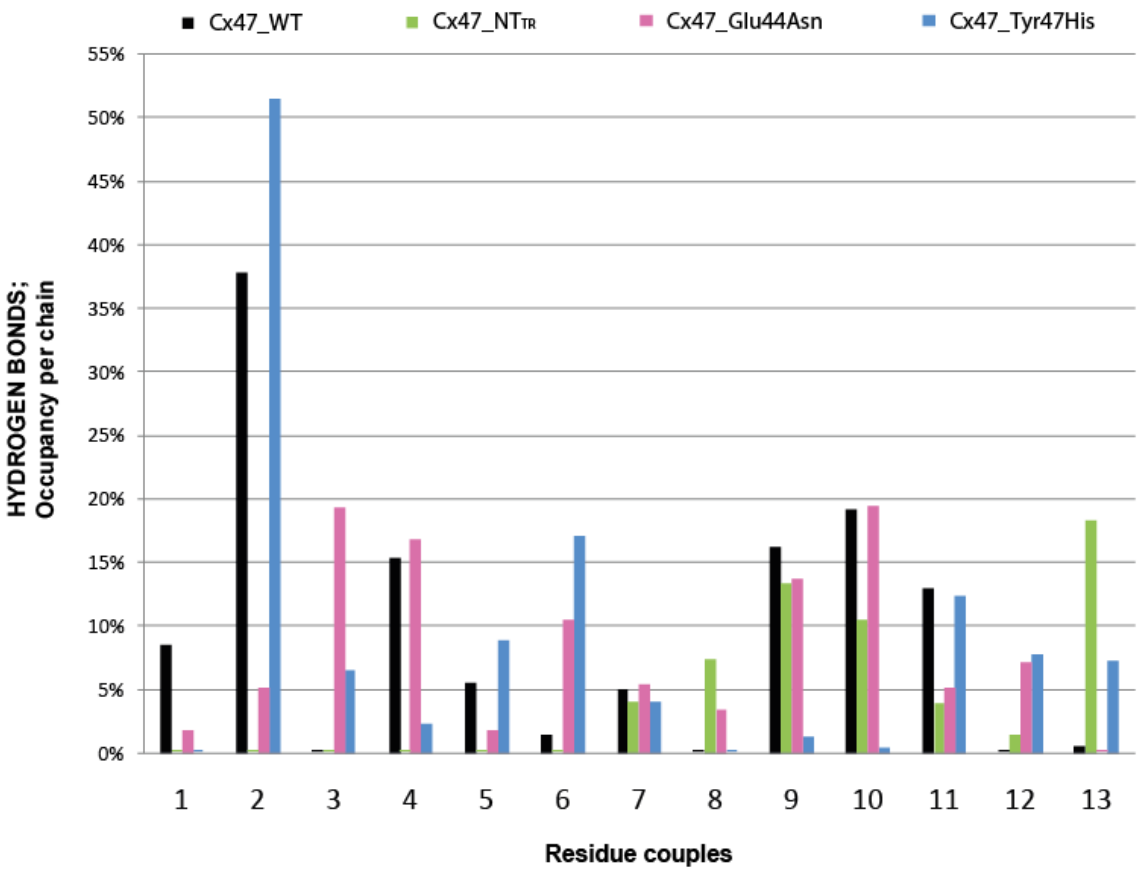


Figure 4-16. *Hemichannel interactions affected by chain modifications.*

The interaction most affected by the truncations is Met1-Glu44 (Table 4-1, row no. 2). In the shorter construct, Cx47_$NT_{TR,}$ such interaction is not present because the NT domain is too short. The Met1-Glu44 interaction seems to be crucial for the NT domain positioning inside the pore and, therefore, for the maintenance of the funnel architecture and its regulation. The neutralization of glutamic acid in the position 44 by asparagine in the full length hemichannel gave similar effect – interaction between residues 1 and 44 significantly decreased. Glu44Asn mutation also altered interaction between residues 44 and 48. Those observations together strongly highlight the importance of the cooperation between Met1, Glu44, Ser48 for a proper (hemi)channel gating (Figure 4-17, yellow ellipses). Another change introduced with the truncation and worth noticing is disappearance of Met1-Met1, Asn3-Asn3 and Asn3-Met4 hydrogen bonds (Figure 4-17, pink ellipses) that stabilize the funnel pore in the WT structure. It has to be emphasized that interactions linked to the NT domain involve both inter- and intra-subunit hydrogen bonds that interconnect the six connexin chains (in the Figure 4-19 shown in the example of A and B chains interactions). Thus, changes in the conformation of NT domain in one of them will also affect to the stability of the NT domains in neighbouring chains, consistent with earlier propositions that gating mechanism is a property of the *entire* hemichannel, rather than an individual chain [141].

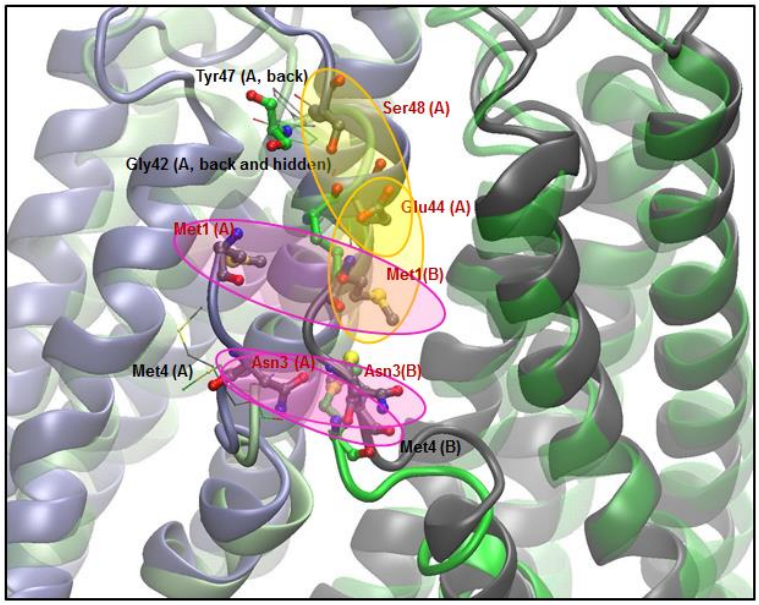


Figure 4-17. Structural effects of NT domain truncation; side view. WT NT domain shown in grey colour, truncated NT domain shown in green; affected or lost interactions marked with yellow and pink ellipses.

In the case of Tyr47His mutation, the most affected interactions are Tyr47-Lys189 and Tyr47-Gly42 (Table 4-1, row no. 9 and 10). Residue 47 is located just on the border between E1 and TM1 domains and is responsible for some intrasubunit contacts within the transmembrane helices in one chain (Figure 4-18).

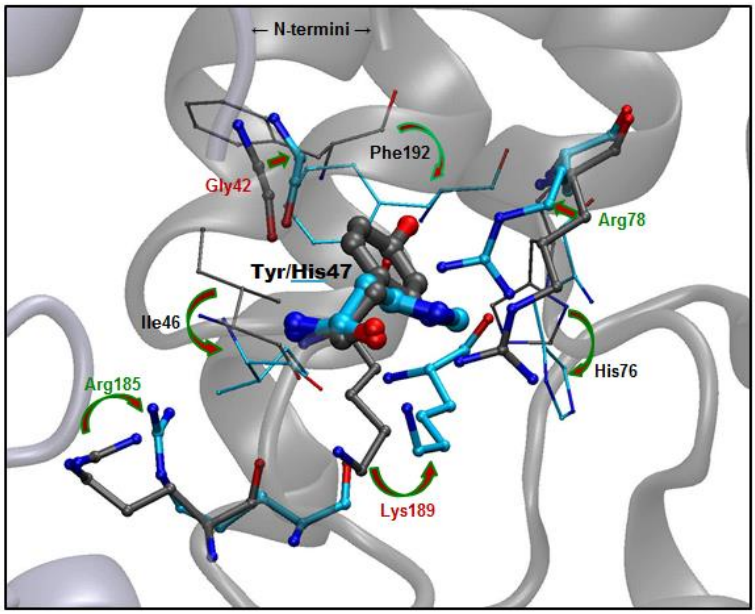


Figure 4-18. Structural effects of the mutation Tyr47His; WT chain shown in grey, mutated chain shown in blue.

It was experimentally proved that Tyr47His mutant is able to form a functional hemichannel but not a full channel, neither with another Tyr47His nor with WT hemichannels. Probably this is due to structural changes affecting extracellular domain, rather than (or more than) pore arrangement. This speculation is even more reasonable when considering the observations already presented in the previous sections: In section 4.1.1.2.1, it is stated that most of the Cx47-Cx47 interfacial interactions being concentrated in the domain E1 and in the section 4.1.2.1 it is shown that the E1

domain is crucial for the channel docking process. Unfortunately visualization of this kind of domain reorganizations fall in a time scale beyond the scope of the existing MD runs.

Figure 4-19 represents the spatial inspection of Cx47 hemichannel in comparison to its altered versions:

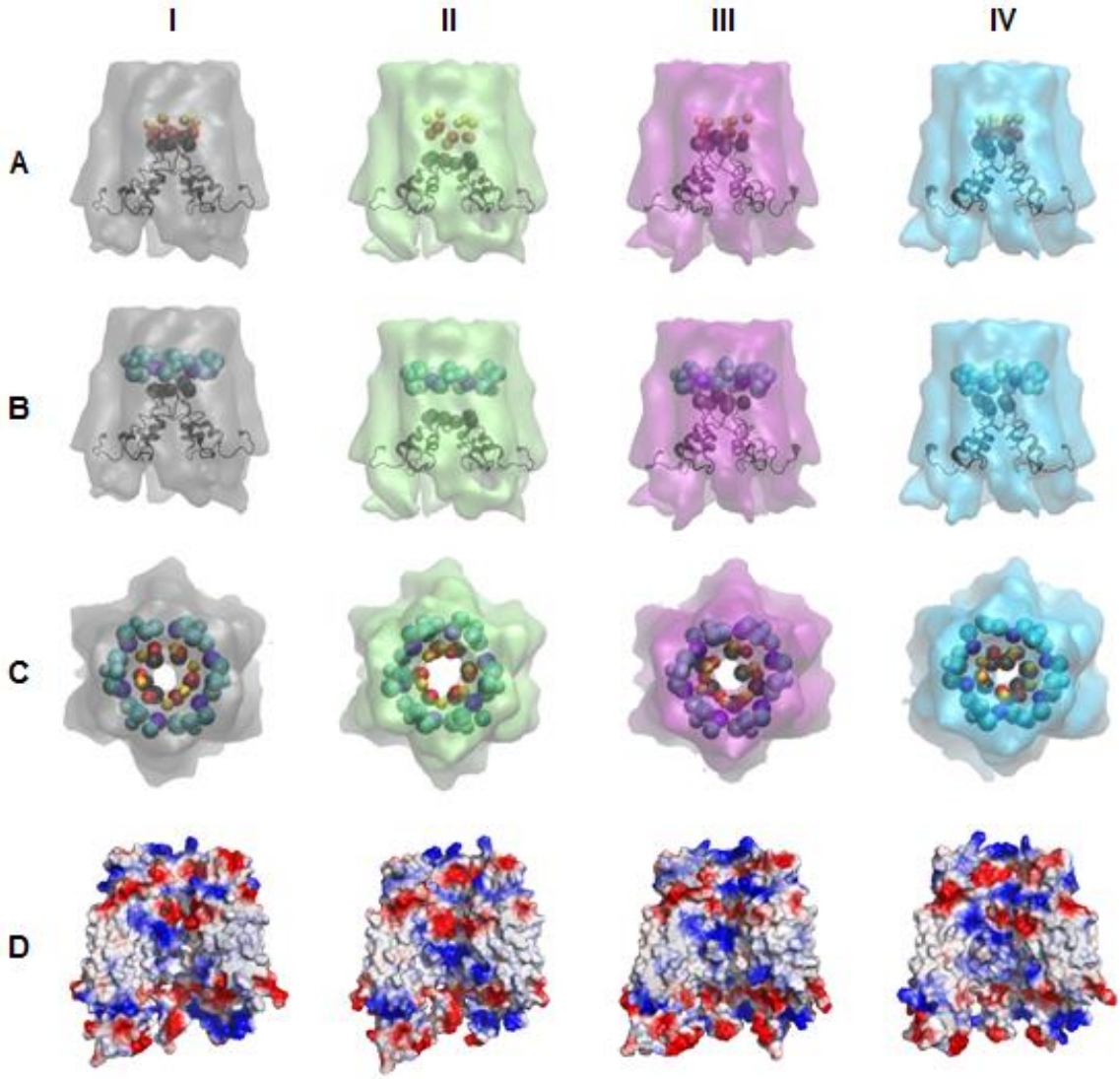


Figure 4-19. Structural effects of Cx47 modifications; general overview.(I) Cx47_WT, (II) Cx47_NT$_{TR}$, (III) Cx47_Glu44Asn, (IV) Cx47_Tyr47His. (A-B) Main residues arrangement, (C) Pore diameter, (D) Electrostatic pore profile. Shown with black cartoon - NT domain, residues 1-20. Shown with beads – Met1 (black), Glu/Asn44 (red), Ser48 (yellow), Tyr/His47 (dark blue), Arg78, Arg185, Glu188, Lys189 (light blue).

Two different groups of interacting residues can be clearly identified. A first group includes Glu44 and Ser48 and they interact with NT domain residues (Figure 4-19A), being affected by the truncation (Figure 4-19A-II) or by Glu44Asn mutation (Figure 4-19A-III). A second group (Gly42, Tyr47, Arg78, Arg185, Glu188, Lys189) scaffolds the structure of a pore (Figure 4-19B) and its organization is affected by both mutations (Figure 4-19B-III, IV). Interestingly, all the modifications determine the pore diameter (Figure 4-19C). NT truncation loosens the funnel and increases permeability while decreases selectivity. Moreover, Tyr47His (shown with details in Figure 4-18) and, to a less extent Glu44Asn, distort the funnel arrangement. Changes in the net charges are not significantly marked in the electrostatic pore profile (Figure 4-19D).

In summary, dynamic changes in interactions and electrostatics correlate with dynamic changes in the structure and, therefore, they ultimately modulate function and stability of a molecular system. All-atom MD simulations of connexin hemichannels results in the reorganization of amino-acid interaction net and present its more realistic model. It allows in-depth analysis of (e.g.) protein organization, regulation and side effects of the introduced modifications, which – if properly planned – shed light on natural mechanisms and/or its disorders. Simulations of the altered versions of Cx47 protein allowed examination of two different hemichannel regions (NT domain and E1/TM1 motive) that are speculated to act as two different gating mechanisms (Vj- and loop-gating). We have shown that both, NT truncation and Cx47_Tyr47His mutation, distort the hemichannel pore affecting its functionality (experimentally proved). NT truncation does it by breaking the intersubunit hydrogen bond network and modifying channel permeability. Cx47_Tyr47His disrupts intrasubunit interactions, causing structural instabilities that may imply difficulties in the docking process.

PERSPECTIVES

Being aware of the limitations of my model (e.g. lack of membranes) and method (rigid docking) I would like to continue the project developing model more similar to the natural environment and conduct simulations of the docked systems to revise and extend my current results.

4:2

HEME PATHWAY ENZYMES
protein-ligand interactions

Two out of eight enzymes of the heme pathway – PBGD and UROIIIS– have been objects of interest of our group for years[45,149–151]. Some mutations in their primary structure cause porphyrias, a group of rare diseases. Till now they are largely incurable and their symptoms (neurological complications and/or skin problems) hamper much of the normal life of patients. The aims of our research are: i) to find and new therapeutic intervention lines based on the use of chemical compounds which may serve as inhibitors (in the case of PBGD) or pharmacological chaperons (in case of UROIIIS), ii) to investigate structures of the enzymes, shedding light on the mechanisms of reaction and inhibition.

4.2.1 Inhibition of PBGD to avoid URO I accumulation.

The PBGD enzyme has been previously subjected to intense mechanistic studies [152]. The identification of the substrate binding site has involved the development of "suicidal" inhibitors, so-called "terminators" that covalently bind to the enzyme and prevent the progression of the reaction – derivatives of the enzymatic substrate such as 2-bromoporphobilinogen [36,153] or 6-methylporphobilinogen and 6,11-ethanoporphobilinogen[154]. It has also been found that porphyrins can inhibit enzyme [155]. Despite their specificity, their efficiency as *in vivo* inhibitors has not been demonstrated, and their high reactivity rules them out as potential drugs. It is also expected for the PBG derivatives to generate many problems *in vivo* since they bind covalently to the enzyme, permanently inactivating it. Furthermore, the substrate is a secondary metabolite and therefore analogous inhibitors could have inhibitory activity in other enzymes reducing the *in vivo* selectivity.[60]

Given that there is a cysteine residue crucial for catalysis (Cys261 in human protein), PBGD modifying agents have also been studied. It has been found that some substances such as metal ions or 2-mercaptoethanol can significantly alter the catalytic activity of the of *A. thaliana* PBGD[156]. However, this is due to a non-specific chemical modification mechanism of the enzyme and not due to a specific binding of the molecule in the active site of the protein. In view of aforementioned, there is still a need of providing therapeutic agents with inhibitory capacity to inhibit the catalytic activity of PBGD potentially useful in the prevention and/or treatment of congenital erythropoietic porphyria (CEP). In addition to effectiveness, compound inhibitors should also not cause undesirable toxicity. Likewise, from their synthesis point of view and pharmacological screening point of view, it would be highly desirable that the compounds can be easily modulated.[60]

4.2.1.1 Theoretical design, docking and selection as a key step in narrowing and directing the search range

The purpose of this study was to analyse by computational docking the inhibitory action of compounds mimicking natural substrate of PBGD and select potentially best candidates for further activity assays. For this, the substrate (PBG) was used as a ligand model and the crystallographic structure of molecular target (human PBGD, wild-type) was obtained from PDB database (PDB ID: 3ECR). For detailed protocol – check section 3.1.4.2 of "Materials and Methods".

Compound selection was based on few criterions in a specific order: (i) substrate similarity determined on 2D text representation (SMILES) string for choosing only the structures most similar to PBG, (ii) water solubility – for reducing potential difficulties during enzymatic assays, (iii) market availability – compounds easily accessible either commercially or by means of conventional synthesis methods from commercial or widely known compounds or intermediates. Ultimately, from the list of ~600 compounds, ~220 were subjected to the further preparation and docked.

For the docking analysis three additional criterions were adopted: (i) active site localization, since the goal is to obtain competitive inhibition, (ii) binding energy value: selecting for the lowest energy value compounds, (iii) interactions – preferably hydrogen bonds – with the residues reported in literature as significant in the reaction process or substrate positioning.

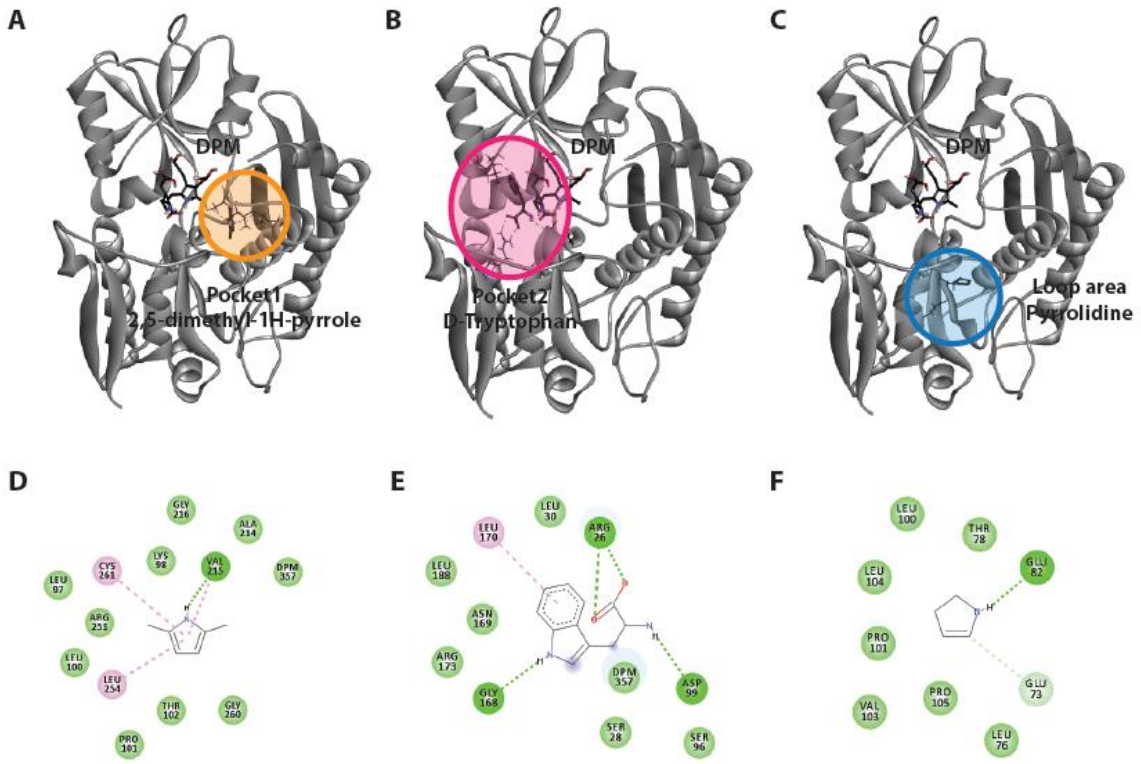


Figure 4-20. Binding sites of PBGD enzyme: Preferable - (A) Pocket 1 represented by 2,5-dimethyl-1H-pyrrole, (B) Pocket 2 represented by D-Tryptophane, Additional – (C) a compound in the loop area, represented by Pyrrolidine. 2D representation of compound interactions with PBGD protein: (light green circle) van der Waals interactions, (green circles and dotted lines) conventional hydrogen bonds, (green/grey) carbon hydrogen bond, (pink circles and dotted lines) pi-pi interactions, (D) 2,5-dimethyl-1H-pyrrole, (E) D-tryptophan, (F) Pyrrolidine.

Two mayor groups of compounds were obtained from the docking analysis. One, denominated later as “Pocket 1” is represented by 2,5-dimethyl-1H-pyrrole (

Figure 4-20A) and gathers small, not branched compounds (pyrrole-based mainly, indole-based occasionally). Second – “Pocket 2” – includes mainly indole-based compounds with variety of substituents. It is represented by D-Tryptophan and shown in the

Figure 4-20B. A few compounds that yielded positive docking results could not be assigned to any of the mentioned groups. As an example Pyrrolidine (

Figure 4-20C), which docked conformation suggest it may affect protein loop mobility. The loop covers the active centre and was proposed to be an important structural motive in the process of substrate accumulation and releasing[34].

Analysis of the intermolecular interactions shows that the selected compounds with the lowest energy complexes are stabilized by intermolecular hydrogen bonds and stacking interactions. Members of the first group are preferably located close to the covalent bond linking DPM cofactor with the conservative Cys261. This gives rises to the idea that these compounds may disturb the correct position of the cofactor in the active centre. Moreover, "Pocket 1" representatives show hydrogen bonds created with the residue Val215 (Figure 4-20D). On the other hand, "Pocket 2" compounds also show specific features: first, they interact with residues involved in substrate positioning (Arg26) and reaction catalysis (Asp99), thus competing with the substrate to bind non-covalently to the active site of the protein (reversible inhibition). Moreover, docked structures are settled in a way that they occupy the incoming substrate position, at the same time blocking the entrance to the active centre (Figure 4-20E).

4.2.1.2 In vitro enzymatic assays for the validation of the hit compounds obtained by molecular docking.

Several assays have been developed to determine the activity of PBGD that were based on UV—vis spectrophotometry[157], fluorimetry[158], or the combination of HPLC and fluorimetry[159]. Product-based direct assays of PBGD however suffer the essential difficulty of having to detect HMB, which is labile and difficult to prepare in a pure state. It spontaneously cyclize to URO I and on to further symmetric metabolites. These properties severely hamper measurements of response and method calibration. The difficulties with HMB can be overcome by including UROIIIS in the assay, the next enzyme in the pathway, and monitoring the production of URO III, which arises by enzymatic cyclization of HMB and is stopped at this stage because of the lack of further pathway enzymes. This kind of indirect, coupled assay – based on the method described by Jordan[108] with some modifications – was used in the present study. Detailed protocol can be found in the section 3.2.2.1.1.

The assay showed good reproducibility and linear dependence of the porphyrin formation (expressed as absorbance changes, ΔOD_{405} and validated by HPLC analysis) on substrate and inhibitor concentration and at the fixed concentration of proteins. The Lineweaver-Burk (LwB) plot depicted in Figure 4-21 shows the inhibition capacity of 2,5-dimethyl-1H-pyrrole ("Pocket 1") and D-tryptophan ("Pocket 2") on PBGD. Both show capacity for inhibiting the catalytic activity of the enzyme following a competitive inhibition scheme wherein the inhibitor and the substrate compete to bind non-covalently to the active site of the protein. The change in slope but with the same intercept in a LwB plot is indicative of a competitive inhibition and thus, confirms the relevance of the molecular docking assumptions and analysis.

The assay tests demonstrated that proposed compounds inhibit the activity of the PBGD since less amount of pre-uroporphyrinogen is produced. This fact may be used as a base for a therapy development against congenital erythropoietic porphyria (CEP). Along with the decreased level of

pre-uroporphyrinogen, the accumulation of URO I derivatives and eventual malfunction of UROIIIS in patients is reduced.

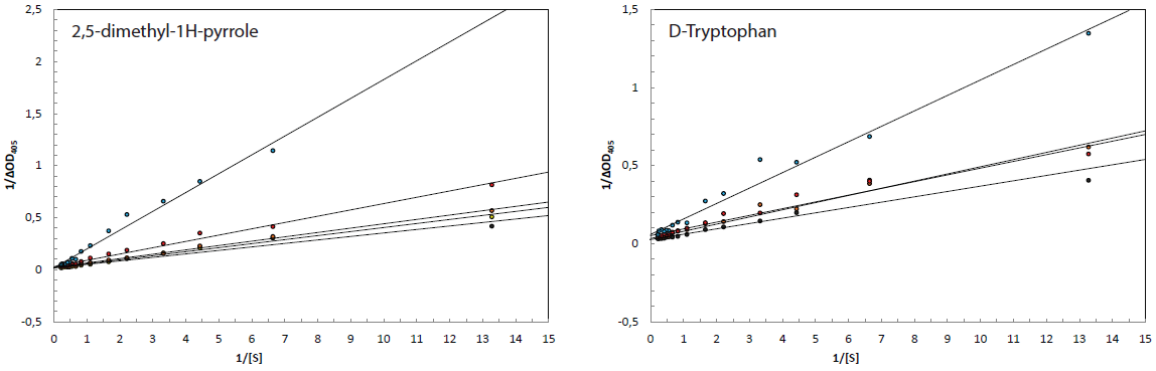


Figure 4-21. Lineweaver-Burk plot for PBGD in the absence (black dots) and in the presence of an increasing amounts of inhibitors. (Left) 2,5-dimethyl-1H-pyrrole: (yellow) 0,025mM, (orange) 0,1mM, (red) 1mM, (purple) 5mM. (Right) D-tryptophan: (orange) 0,5mM, (red) 1mM, (purple) 5mM. The symbols correspond to the experimentally determined values meanwhile the lines show the linear regressions thereof.

Tested compounds constitute a new line of inhibitors[60] based on reversible inhibition. Remarkably, they have an “inhibiting” moiety, a main part responsible for the activity, and a “modular” moiety, which may allow tuning the pharmacological properties of the compound without significantly altering the inhibitory capacity of the molecule. Furthermore, since some of these compounds are natural metabolites or chemically resemble to natural metabolites, it is expected that for the toxicity of the compounds to be low. This fact is especially important in the case of rare diseases such as in the case of CEP, since the lack of patients limits or makes the development of the clinical phases of the medicinal product practically impossible. The chemical structures of the aforementioned compounds are presented below:

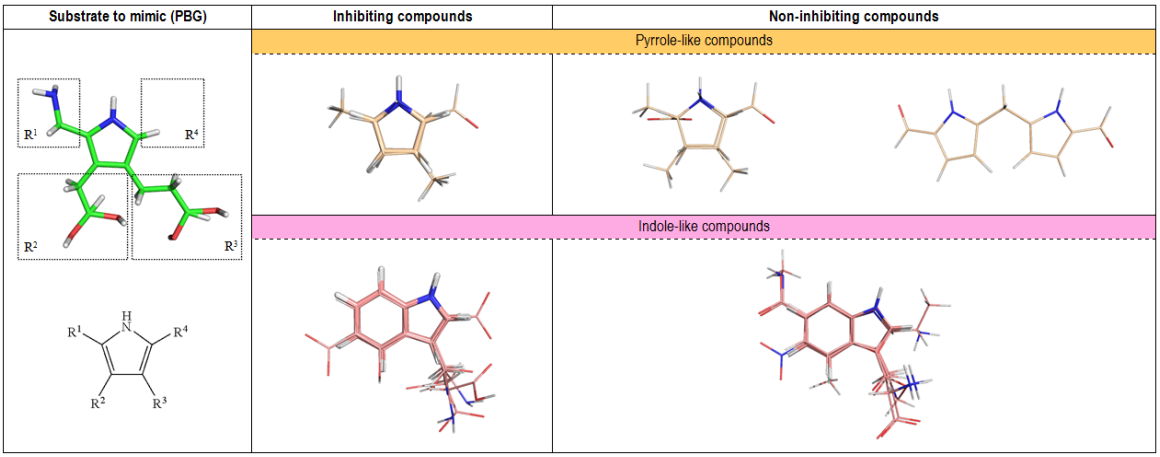


Figure 4-22. Chemical structures of tested compounds. (Left, top) Structure of PBG – natural substrate, (left, bottom) schematic representation of investigated structures, (centre) inhibiting compound, (right) non-inhibiting compounds, (orange) pyrrole-like compounds represented e.g. by 2,5-dimethyl-1H-pyrrole, (pink) indole-like compounds represented by e.g. D-tryptophan.

All the tested compounds are schematically presented in a form shown in the left-bottom part of the Figure 4-22.[60] R^1 and R^2 can be H and C_1-C_6 alkyl, or, alternatively, R^1 and R^2 are bound to one another forming an optionally substituted fused benzene ring. R^3 is can be H, C_1-C_6 alkyl or other

functional groups like amine (-NH_2) and carboxylic acid (-COOH). R^4 can be H or C_1-C_6 alkyl. "Alkyl" refers here to a saturated linear or branched hydrocarbon chain radical consisting of carbon and hydrogen atoms which are bonded to the rest of the molecule by means of a single bond. When "C_1-C_6 alkyl" is indicated, it refers to a alkyl group of one, two, three, four, five or six carbon atoms such as for example, methyl, ethyl, propyl, isopropyl or n-butyl.. Furthermore, preferred compounds are those wherein at least two of R^1, R^2, R^3 and R^4 simultaneously represent methyl. However, not every multi-branched structure is accepted (Figure 4-22, right). On the other hand, Table 4-2 represents juxtaposition of particularly preferred compounds, arranged from best to worse inhibitor.

INHIBITOR	Computational procedures				Enzymatic Assays				
	Binding energy (ΔG) kcal/mol	Inhibition constant (Ki) μM	WS mM	MW g/mol	[S] reaction solution mM	[I] eaction solution mM	$ABS_{W/inh}$	$ABS_{W/Oinh}$	INHIBITION %
2,5-dimethyl-1H-pyrrole (A008)	-3,84	1530	263.03	95.14	0,60	1,80	0,7307	1,9413	62,36
C009	-5,79	57,24	3.98	189.21	0,60	3,82	0,1077	0,5133	79,02
C003	-5,33	123,96	8.51	161.16	0,60	5,78	0,0038	0,5450	99,30
C021	-5,48	96,33	9.33	161.16	0,60	5,78	0,0049	0,4730	98,96
A007	-3,95	1260	263.03	95.14	0,60	5,78	0,0076	0,4097	98,14
C020	-4,33	667,04	29.51	133.15	0,60	5,78	0,0376	0,7218	94,79
A001	-2,96	6810	1230.27	67.09	0,60	5,78	0,0825	0,4774	82,72
C011	-4,66	381,85	7.59	204.23	0,60	5,78	0,1568	0,3833	59,09
C001	-4,24	774,31	18.20	117.15	0,60	5,78	0,2130	0,4774	55,38
C008	-5,32	125,77	10.23	161.16	0,60	1,45	0,4790	0,5450	12,11
D-tryptophan (C010)	-4,56	454,15	7.59	204.23	0,60	5,78	0,2171	0,3833	43,36
C002	-3,85	1500	35.48	119.16	0,60	5,78	0,3380	0,5792	41,64
Pyrrolidine (A005)	-3,58	2,38	758.58	71.12	0,60	5,78	0,4058	0,5133	20,94
A009	-5,43	103,8	104.71	123.15	0,60	5,78	0,4720	0,5792	18,51
C017	-4,16	890,67	2.29	156.18	0,60	5,78	0,4488	0,5133	12,57

MW molecular weight, WS water solubility, ABS absorbance, [I] inhibitor concentration, [S] substrate concentration

PBG MW: 226.23 g/mol

Table 4-2. Summary of PBGD inhibitors; 15 best cases. (Orange) Pyrrole-like compounds represented e.g. by 2,5-dimethyl-1H-pyrrole, (pink) indole-like compounds represented by e.g. D-tryptophan. Full names replaced by a code system to protect the undergoing patent registration procedure.

According to the assays, the most promising compound is 2,5-dimethyl-1H-pyrrole which demonstrated the highest level of inhibition in the relation to its concentration in the reaction solution. Thus, it has been established as a reference for further trials. Most of the listed compounds represent bigger, pyrrole-like structures however, they all share the fact that the molecular weight does not exceed the molecular weight of the natural substrate. This is a feature which may promote them in the competition for the active centre with PBG.

Table 4-2 presents also a summary comparison between computational (left) and experimental (right) data. It is clear that neither binding energy, nor inhibition constant estimated during the docking procedure can precisely predict real inhibition level. It does not change the fact that computational approach has correctly identified a group of competitive inhibitors from which hit molecules were selected and now are serving as a fundament for CEP therapy development. Analysis of Table 4-2 reveals that the most effective way to characterize protein-ligand interactions is by combining both, computational and experimental data.

These results indicate that PBGD can be effectively inhibited by small organic molecules so a massive screenings of compounds libraries and *in vivo* tests that are being currently conducted in the laboratory. Ideally, in case that these compounds reach further stages of the drug development, the active principle will have to be formulated in an acceptable or substantially pure form, in a

pharmaceutically suitable composition and in a therapeutically effective amount together with one or more carriers, adjuvants or excipients for its administration (e.g. sterile liquids such as water and oils) by any suitable administration route (e.g. oral or parenteral). In theory, the treatment could/should be complimented with the administration of heme derivatives or blood products to counteract the heme group deficiency.[60]
Success and advancement of the activity studies also encouraged us to go a step further and try to characterize PBGD mechanism. The attempts are presented in the following sections.

4.2.2 PBGD mechanism

The most intriguing feature of PBGD is the reaction it catalyses: PBGD forms linear tetrapyrrole from four units of porphobilinogen using a single active site, by repeating the same action four times. The existence of the intermediates were proven before (works of Jordan PM and collaborators) but never resolved at the structural level. The structural determination of the reaction intermediates would finally reveal the speculated mechanism facilitating protein-substrate/ligand investigation and thereby e.g. an inhibitor and potential porphyria therapy development.
To that end, we have optimized protocols for expressing and purifying wild type and mutant PBGDs (also labelled with ^{15}N isotopes for NMR studies), obtaining abundant yield and undergoing the separation of the intermediates of the reaction. This protocol allows the preparation of pure and highly concentrated samples of the intermediates, later on used for crystallization, SAXS and NMR experiments. Our results prove the dynamic character of the enzyme and agree with the previous predictions, although, the quality of data (in case of NMR) and the resolution (in case of SAXS and crystallography) still are not enough to explore atomic details.

4.2.2.1 Purification and reaction intermediates separation

To obtain samples of the PBGD reaction intermediates we performed a basic purification process as previously published [153,160–162] and described in the section 3.2.1.2 of “Materials and Methods”. This basic procedure is sufficient for obtaining large amount of pure sample (mixture of the intermediates) used e.g. for enzymatic assays.
The major problem to isolate the individual ES complexes of PBGD is related to the inherent reversibility of the substrate complex[153] - under some conditions the substrate can levy and attach itself to another ES complex - thus randomising the isolated form. Isoform separation was achieved using ion exchange chromatography (see below) and, as long as a sample is not exposed to harsh chemical conditions this randomization can be largely controlled. Each complex, once separated, is rather stable since re-injection in the cation exchange column did not separate again into several peaks (Figure 4-23B). Moreover, samples can undergo (re)freezing without significant losses of activity.
To the separation protocol were subjected following samples: wild-type human housekeeping PBGD protein (PBGD1) and its truncated version, corresponding to the wild-type erythrocytic enzyme (PBGD2) - Figure 4-23. Also a mutated protein – PBGD1_R26V (Figure 4-24) and

isotopically labelled proteins – ^{15}N-PBGD1 and ^{15}N-PBGD1_R26V (Figure 4-25). The mutation R26V was chosen with the intention to affect the active centre ("Pocket 2" like so) and thus observe and characterize eventual changes in the interaction net. Arg26 was reported to be among the residues involved in the correct positioning of incoming molecules[34].

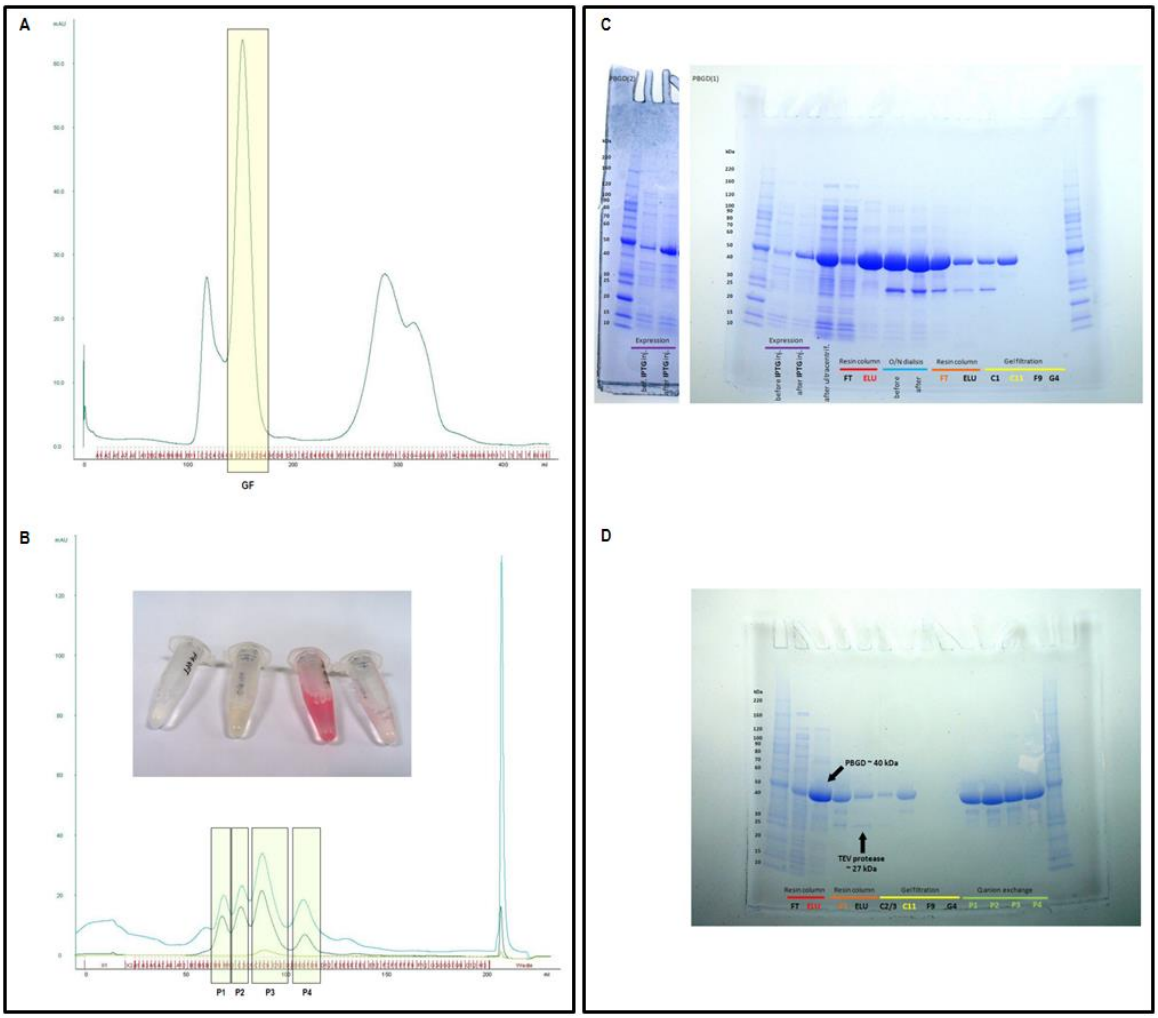


Figure 4-23. Wild-type PBGD purification. (A) Size exclusion chromatography (GF, gel filtration) – elution profile of wt PBGD1 in *HiLoad 26 600 Superdex75*, blue line corresponds to absorbance in arbitrary units (vertical axis), horizontal axis shows the elution volume and the elution fraction. (B) Anion exchange chromatography of wt PBGD1 in HiTrap Q HP 5ml (Q) with samples corresponding to the peaks of the successive reaction stages presented in the picture, blue line – anion exchange chromatography with size exclusion chromatography step excluded from the protocol, dark green line – anion exchange chromatography proceeded by size exclusion chromatography step, light green line – double anion exchange step: Q re-injection of the P3 peak. (C) SDS-PAGE gel for wt PBGD1 purification process from expression to size exclusion chromatography (GF) step; (left) PBGD2 expression as a proof of construct behaviour similarity. (D) SDS-PAGE gel for wt PBGD1 purification process from immobilized metal ion affinity chromatography (Ni-column) to anion exchange chromatography (Q) step.

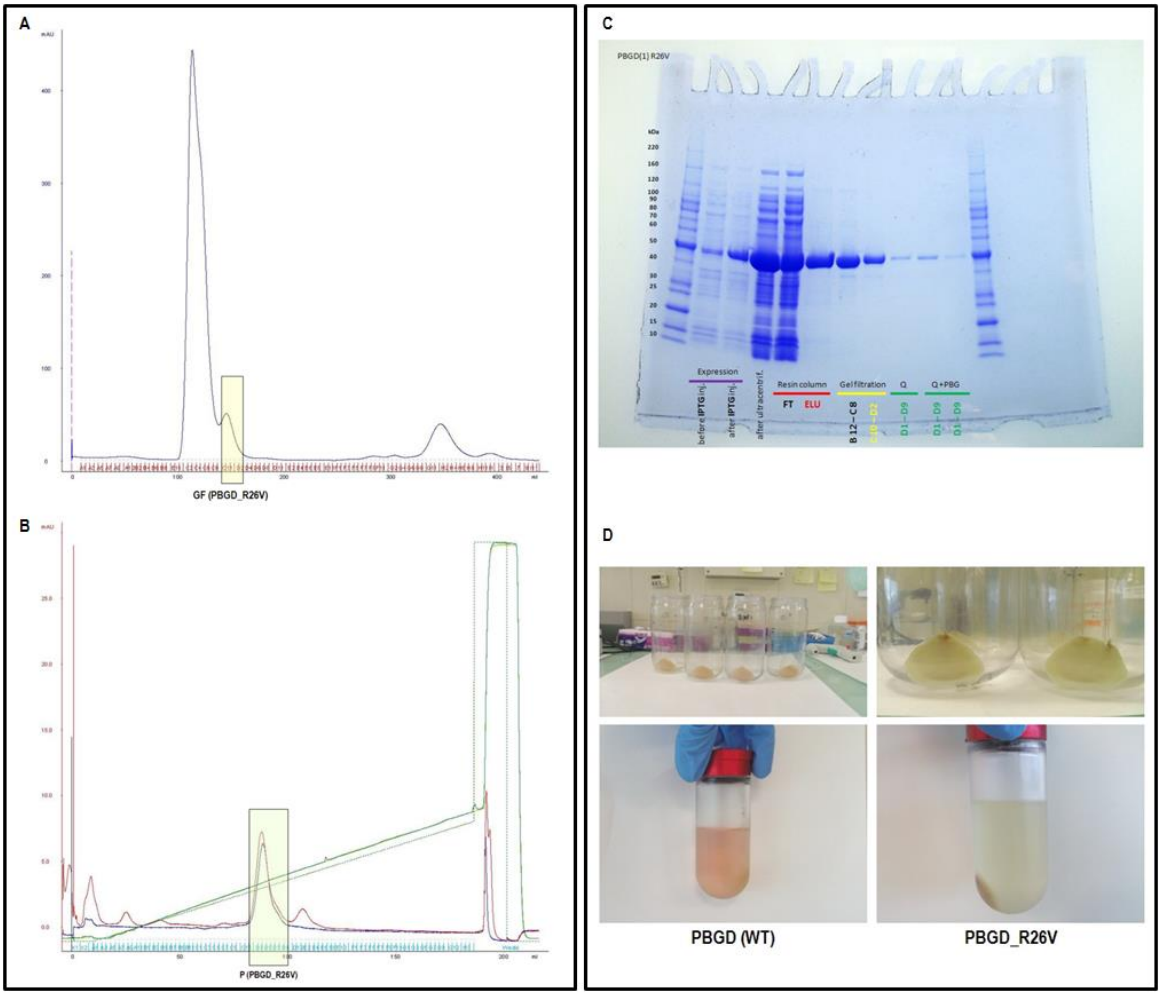


Figure 4-24. PBGD1_R26V purification. (A) Size exclusion chromatography – elution profile of PBGD_R26V in *HiLoad 26 600 Superdex75*, blue line corresponds to absorbance in arbitrary units (vertical axis), horizontal axis shows the elution volume and the elution fraction. (B) Anion exchange chromatography of PBGD1_R26V in HiTrap Q HP 5ml. (C) SDS-PAGE gel for PBGD_R26V purification. (D) Colour differences between wild type (left) and R26V mutant (right) protein cultures and supernatants.

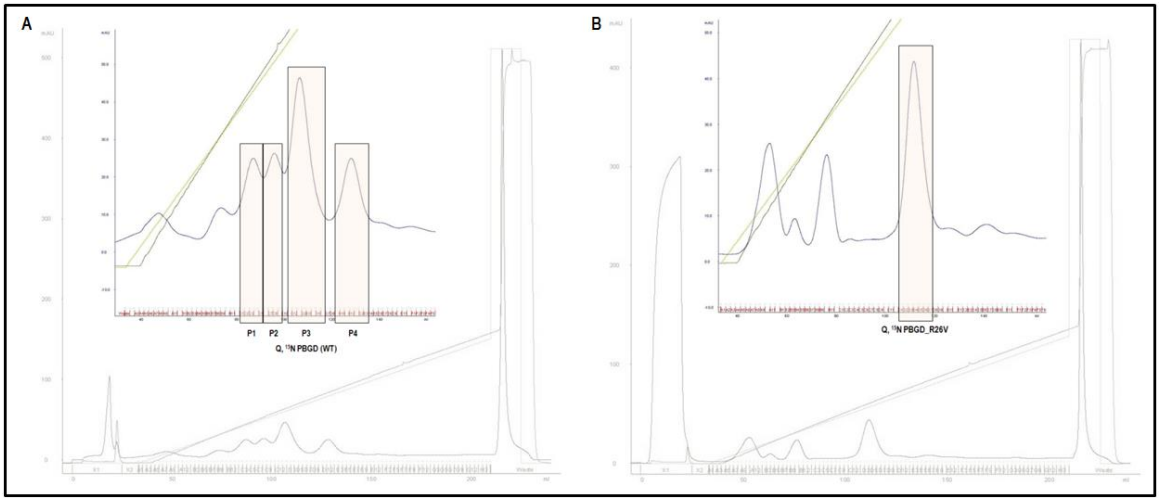


Figure 4-25. Anion exchange chromatography of ^{15}N isotopically enriched protein in HiTrap Q HP 5ml. (A) ^{15}N-PBGD1: (bottom) full chromatogram, (top) zoomed chromatogram. (B) ^{15}N-PBGD_R26V: (bottom) full chromatogram, (top) zoomed chromatogram.

Both, His-tag and imidazole concentration prevent protein binding to the anion exchange columns and they have to be eliminated. Moreover, long histidine tail can effectively disturb protein crystallization. To that end, we used homemade His-tagged TEV protease with S219N mutation that removes a self-cleavage site for PBGD1 and PBGD1_R26V purification (protease concentration 1/20 with respect to sample concentration) and commercially obtained Factor Xa protease (New England BioLabs) for PBGD2 purification (protease concentration 1/20 with respect to sample concentration). In both cases dialysis undergoes at ~18 ºC.

Size exclusion chromatography (gel filtration) step, although usually applied as the final purification step, here precedes anion exchange chromatography. Sample washed out from the Ni-column still contained unwanted contaminants thus, swapping the column order turned out to be useful to avoid column blocking, column overloading or eluting additional peaks (Figure 4-23B, compare dark and light green lines). This is particularly important, considering the precise and delicate MonoQ column. Another important problem is that anion exchange chromatography profile differs depending on the conditions and/or the constructs, PBGD1 and 17 aa-shorter PBGD2, respectively. In the Figure 4-23B Q column chromatogram is presented and is also compared to MonoQ chromatograms in the supplementary figures Sup. Figure 6-1 (PBGD1) and Sup. Figure 6-2 (PBGD2). Protein purity is further characterized by SDS-PAGE and Native gel electrophoresis, along with mass spectrometry analysis (mass-spec).

Mass-spec confirms the separation of the PBGD isoforms. At the same time it revealed a certain degree of contamination between isoforms that required further protocol optimization. The problem was alleviated after exchanging Q for MonoQ column. It has to be taken into consideration that mass spectrometry is not a quantitative method and mass spectra for purified and re-purified samples can be found and compared in the supplementary materials (Sup. Figure 6-3 – Sup. Figure 6-7 for peaks P1 – P4, respectively). The isoforms were further characterized by electrophoresis techniques, SDS-PAGE and native gel (carried under commonly available protocols and easy to adapt), which allows quite precise protein monitoring.

Denaturing gels (usually SDS-PAGE) are run under conditions that disrupt the natural structure of the analyte, causing it to unfold into a linear chain. Thus, the mobility of each macromolecule depends only on its linear length and its mass-to-charge ratio. Thus, the secondary, tertiary, and quaternary levels of biomolecular structure are disrupted, leaving only the primary structure to be analysed.[4] This technique was useful to track PBGD sample in process of purification. Examples can be seen in Figure 4-23C-D and Figure 4-24C.

On the other hand, native gels (native PAGE) are run in non-denaturing conditions, so that the analyte's natural structure is maintained. This means that the physical size of the folded or assembled complex dominates the mobility, allowing for analysis of all four levels of the biomolecular structure. Complexes remain—for the most part—folded as they would be in the cell, differing in molecular mass and intrinsic charge.[4] This quality has proven to be very useful in PBGD complex monitoring and has been used to sample homogeneity determination. Figure 4-26 presents native gels run for all PBGD constructs where GF_{PBGD1} sample (first row) served as a ladder.

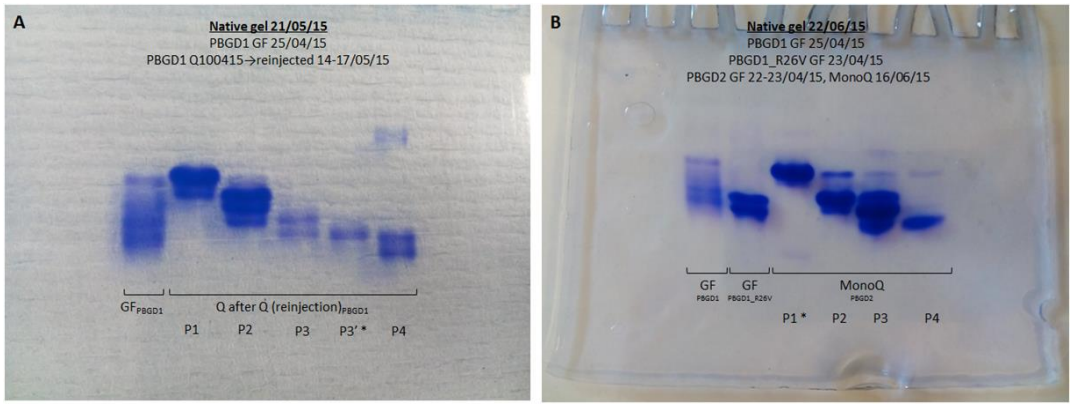


Figure 4-26. Native gel electrophoresis: (A) PBGD1 peaks composition, (B) PBGD1_R26V and PBGD2 samples composition. Marked with asterisk (*) – samples successfully crystalized.

Combining mass-spec and native gel techniques PBGD peaks were determined as follow:

- P1 » E_{HOLO} as a major part of a sample
- P2 » ES_1
- P3 (eventually also P3') » ES_2
- P4 » ES_3
  - ES_4 complex has not been observed

This sorting contradicts to the previous assumptions (P1 » ES_1, P2 » ES_2, P3 » ES_3, P4 » ES_4). The combination of sample SDS- and native PAGE is powerful to clarify the isoform composition as shown in the example of PBGD1_R26V (Figure 4-27), where SDS-PAGE method (left), along with size exclusion chromatogram, was used for selecting fraction containing most appropriate fraction which then was loaded to anion exchange column. Each eluted fraction was further analysed by native PAGE and only those exhibiting one band (or near one band) they were subjected to the crystallization studies where – to obtain single, good diffracting crystals – samples have to be very pure and homogenous. Consequently, the entire purification procedure results in the loss of a large amount of sample and the final yields are moderate.

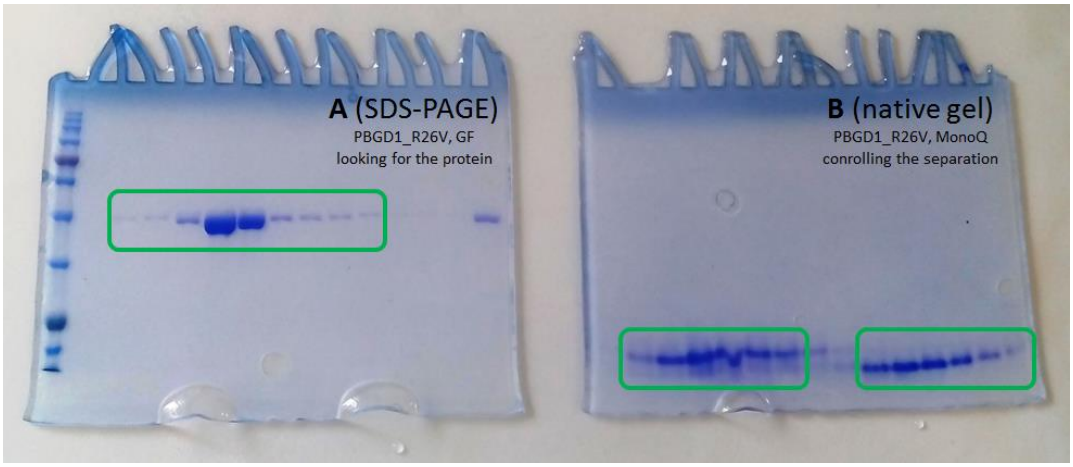


Figure 4-27. PBGD1_R26V gel electrophoresis: (A) SDS-PAGE after the step of size exclusion chromatography, (B) native gel after the step of anion exchange chromatography, (green) fractions taken for further analysis.

Unlike many other proteins the PBGD enzymes from most sources (also mutant proteins which folding is not affected) exhibit reasonable thermal stability and thus, so the purification process does not estrictly require working in strict 4 ºC. However, it is recommended sample should to be kept at low temperature as much as possible to minimize proteolysis. In addition, light sensitivity and pink colour are two characteristic features for PBGDs, both related to the DPM cofactor. It has been suggested that the pink colour of the freshly purified protein is owing to the presence of its fully reduced form[163]. To prevent its oxidation and thus, loose of function longer purification steps (e.g. Ni-column incubation, O/N dialysis) and crystal growing should be carried in dark. Still, over a period of weeks the protein turns yellow, presumably owing to slow but natural and inevitable oxidation of the cofactor.

We have undergone the purification of the mutant R26V using the same protocol. Interestingly, PBGD_R26V mutant is almost colourless (Figure 4-24D). It has shown no activity in the enzymatic assay, also it precipitates significantly (Figure 4-24A) and shows almost no separation in the anion exchange chromatography (Figure 4-24B). This suggests that the mutant may lack the cofactor, based on a previously report that links apoenzyme inactivity and unfolding[161,164]. However, this is unlikely according to the gel filtration profile: it is eluted in the range similar to wt protein and to native gel analysis: Figure 4-26B shows two mutant protein bands at the level of two last bands of wt protein. It suggest that R26V, similarly to previously reported cases (equivalent mutation R11H affects E_{HOLO} » ES_1 stage[161] in *E. coli* and R26A mutation decrease enzyme activity (1.1 %) in humans[34]) – most probably blocks the enzyme in one of reaction steps. Therefor R26V mutation not only may be useful in the reaction intermediate structural characterization trials but – as the R26 was suggested by docking analysis to be crucial in the PBGD-substrate/ligand interaction - it serves as another verification of molecular docking method and proper interpretation and confirmation for its proper interpretation.

Samples prepared as above were then tested structurally.

4.2.2.2 SAXS experiments as the first successful attempts for structural characterization of the PBGD intermediates.

In order to confirm that there is any structural differences between PBGD reaction intermediates or mutant protein we performed small-angle X-ray scattering (SAXS). Data were collected at two different concentrations (for detail check the protocol in the section 3.2.2.2.1) and ultimately analysed for the lower concentration of all samples.

Human PBGD structure was modelled for the missing coordinates (1-20, 60-78, 361-364) using existing templates from PDB (PDB ID: 3ECR) and subjected to the MD simulations above extended N-terminal conformation for structure relaxation (17ns) according to the protocols from sections 3.1.1.1 and 3.1.3. Appropriate cofactors as per the catalytic stage were not included.

Below are presented the computational efforts at fitting the simulated models to SAXS data. The experimental spectra shown along with simulated fits obtained from the model can be find in the supplementary materials (Sup. Figure 6-8 – Sup. Figure 6-12). 3D representatives are shown in the Figure 4-28 and the entire qualitative summary in the Table 4-3.

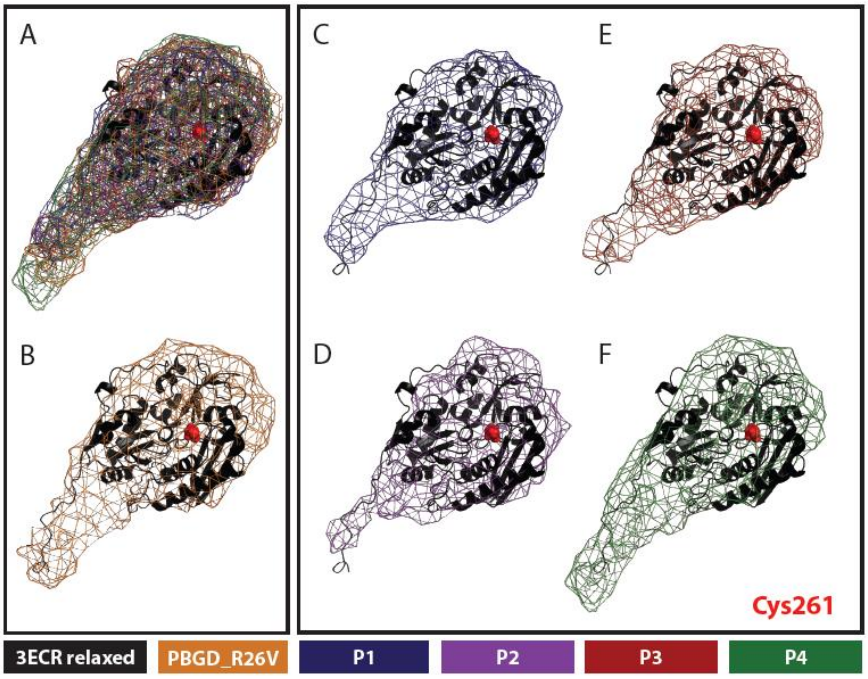


Figure 4-28. Superimposition of relaxed model structure with the respective peak files.

Protein sample	model	frames	Rg [Å]	D_{max} [Å]	Chi	NSD
PBGD1 P1	3ecr_17ns	19-699	2.39 +/- 0.01	8.38	1.094	0.8212
PBGD1 P2	3ecr_17ns	41-682	2.45 +/- 0.07	8.59	2.04	0.8169
PBGD1 P3	3ecr_17ns	42-689	2.43 +/- 0.04	8.51	1.788	0.8117
PBGD1 P4	3ecr_17ns	41-685	2.92 +/- 0.12	10.23	5.037	0.8114
PBGD1_R26V	3ecr_17ns	44-636	2.63 +/- 0.02	9.20	5.525	0.8087

Table 4-3. SAXS data analysis statistics.

The radius of gyration, R_g was estimated from the low-angle scattering region by a Guinier plot[117]. Alignment of the model (3ECR N-terminal extended structure) with the respective peak files was done using Supcomb[120] and the *ab initio* shape reconstruction by DAMMIN[119] using P1 symmetry showed a good agreement between the predicted molecular shape (colourful mesh) and the model (black cartoon); for the quality of the last one, check NSD in the Table 4-3.

Samples differ mainly in the values of maximum linear dimension D_{msx} and Chi. We interpret the higher chi value as arising in part from the active centre area rearrangements to receive and accommodate incoming substrates. However, the changes are not as significant, as suspected. D_{max} value increase mainly in case of PBGD1 P4 – last stage of reaction before product release, and the mutant – according to the native gel analysis (Figure 4-26), stopped at one of the last reaction stages. Those results agree with the NMR analysis (Sup. Figure 6-13) which revealed visible, although modest special rearrangements.

SAXS (same as gel filtration chromatography) analysis additionally confirmed also PBGD to exist as a monomer in solution.

Presented results proved the dynamic character of the enzyme although, the quality of data (in case of NMR) and the resolution (in case of SAXS and crystallography) still were not enough to explore atomic details.

4.2.2.3 Crystallization and preliminary X-ray characterization of PBGD reaction intermediates.

As previously mentioned, to date, ten crystal structures of the protein are available in the Protein Data Bank (2 of *H. sapiens*, of which 1 of wild type protein, 5 of *E. coli*, 2 of *B. megaterium* and 1 of *A. thaliana*). However, the structures of subsequent catalytic stages of PBGD (isoforms), essential to understand the mechanism of the reaction and to characterize the protein-substrate/inhibitor interactions remain unknown. To maximize chances for obtaining diffraction-quality PBGD crystals apart of the purification protocol optimisation (discussed in the previous section) we have explored new crystallization conditions. Different crystallization screenings were tested, based on the conditions reported in literature or using commercially available screens. They were set manually or performed automatically with the Mosquito robot (TTP LabTech Ltd.). Unfortunately, most of the conditions where crystals appeared contained ammonium sulphate and it was suggested by Roberts at al.[165] that any enzyme-bound intermediate can react with ammonia to give PBG and with hydroxylamine or water (at elevated temperature) to give hydroxyporphobilinogen, and both of these products can act as substrates in the condensation reaction. During elongation, random movements of the protein domains may then cause the bound polypyrrole to undergo a random walk, perhaps shuffling backwards and forwards within the active-site cleft in a stepwise manner.
We have further investigated the effect of ammonium sulphate. Three samples (1 – protein sample prepared for crystallization with the standard protocol, 2 – sample 1 diluted 2 times in water, 3 – sample 1 diluted2 times in 1.2 M ammonium sulphate) were incubated 20-30 minutes in the room temperature and then subjected to the native gel electrophoresis. Test was carried for the mutant protein (PBGD1_R26V) and one example of the separated reaction intermediates (P2, PBGD2). As shown in Figure 4-29, the sample with water dilution (= protein concentration change) shows no effect in the isoform composition neither in case of the mutant protein nor the reaction intermediate (2a, 2b). In contrast, ammonium sulphate has caused protein re-separation (favouring P1 – presumable E_{HOLO}) in case of the wild type protein (3b) but not of the mutant (3a).

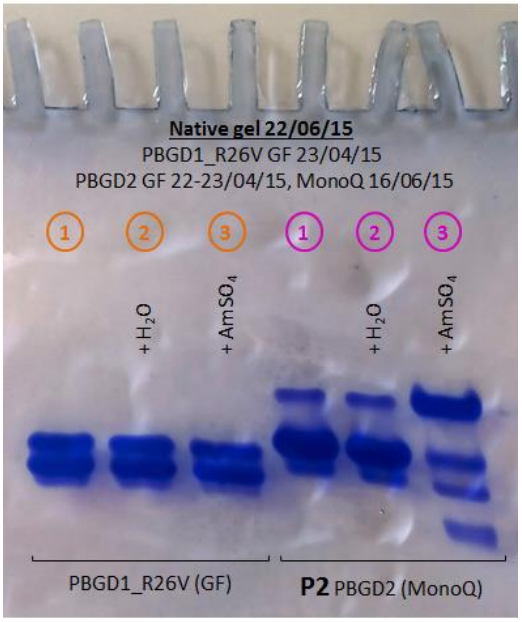


Figure 4-29. Native gel electrophoresis for analysing ammonium sulphate influence on PBGD separation.

It is also worth mentioning that the proposed active site of PBGD is occupied by an acetate ion in the *E. coli* and *A. thaliana* PBGD structures (PDB ID: 1PDA, 4HTG, respectively) and a sulphate ion in the human PBGD structure (PBD ID: 3EQ1). The sulphate ion is located in hydrogen-bonding distance to an arginine and a serine, both proposed to be involved in substrate binding. Summarizing, solutions containing ammonium sulphate may be applied in PBGD holoenzyme (P1 peak) and R26V mutant crystallization trials, but not other PBGD complexes (P2, P3 and P4).

In the course of our studies, over 160 crystals have been sent to different synchrotron centres. Diffraction data collection was only possible for a few samples, in most cases with unsatisfactory resolution range, thus highlighting the enormous complexity of the project. Moreover, PBGD tends to crystalize in the form of multiple plates, that grow tightly one on the other (for examples check Sup. Figure 6-14). Attempts to control the number of nucleating sites by changing the protein and/or precipitant concentration or by adding seeding crystals were rather ineffective. Different temperature settings (4 °C, 10 °C, 19 °C, 21 °C, 25 °C, 37 °C) were also tested. Crystals, to prevent cofactor and substrate oxidation, were always grown in dark.

From among 40 crystals sent to ESRF synchrotron (Grenoble, France), 27 were tested and full data set was recorded (remotely) for 8. Two sets have been successfully processed – for two reaction intermediates (p3', PBGD1, Figure 4-30) and (P1, PBGD2, Figure 4-31).

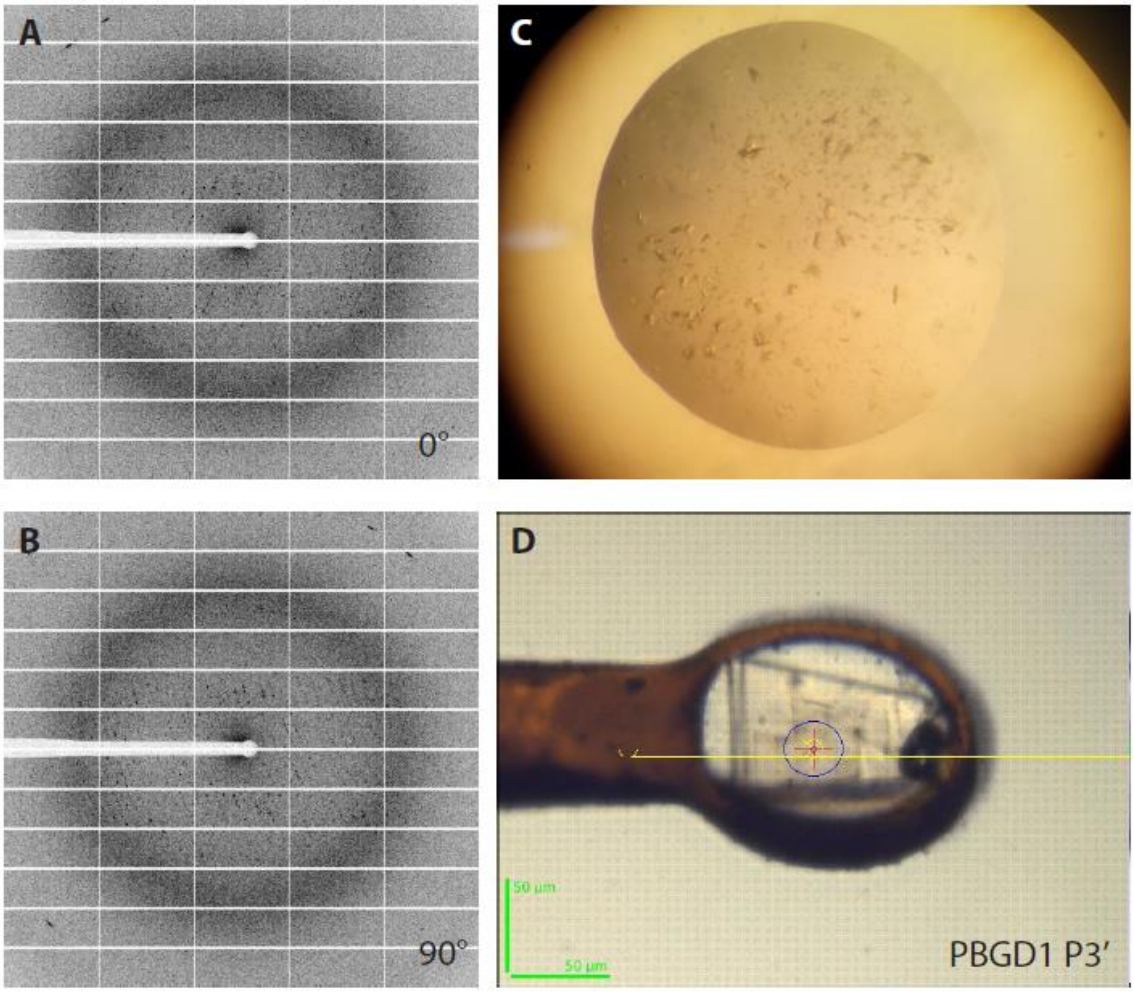


Figure 4-30. (A-B) Diffraction images of PBGD1 P3'. Crystals of PBGD1 P3' in a drop (C) and in a loop (D).

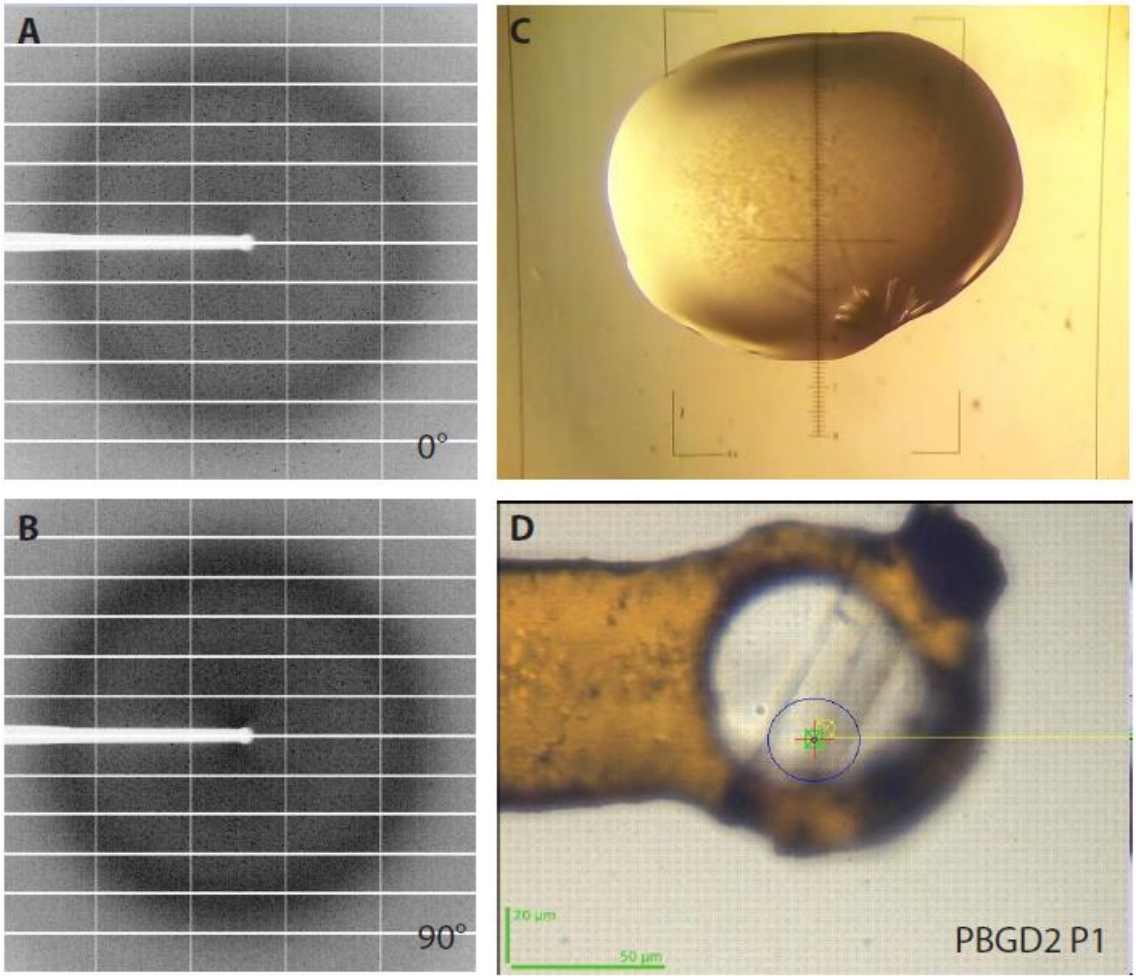


Figure 4-31. (A-B) Diffraction images of PBGD2 P1. Crystals of PBGD2 P1 in a drop (C) and in a loop (D).

Diffraction-quality crystals for PBGD1 P3' sample and PBGD2 P1 sample were obtained by hanging-drop vapour diffusion at 21°C (dark) in a solution containing 0.1 M Na HEPES pH 7.2 (buf.), 0.9 M Na dihyhydrogen phosphate/0.9 M K dihydrogen phosphate (salt) and 0.1 M sodium cacodylate pH 6.5 (buf.), 0.2 M sodium chloride (salt), 2 M ammonium sulphate (salt), respectively. Individual crystals were cry protected by a brief soak in well buffer supplemented with 25% glycerol and flash frozen in liquid nitrogen. More detailed information can be found in the supplementary materials (Sup. Table 6-1).

Native gel imaging PBGD1 P3' and PBGD2 P1 (inter alia) is shown in the Figure 4-26 proving that accurate separation significantly increases the chance of obtaining well diffracted crystals; crystalized samples marked with asterisks. Moreover, as mass-spec results suggests that PBGD1 P3' complex should correspond to ES2 while for PBGD2 P1 should be the E_{HOLO} isoform.

Pinkish (or orangish) crystals appear normally after few days and over a period of several weeks they became darker, presumably owing to slow oxidation. In case of conditions containing salt mixture of Na dihyhydrogen phosphate and K dihydrogen phosphate this process is observed after hours. PBGD1 P3' crystal was fished next day after plate setting. Crystals shown in the Figure 4-30D and Figure 4-31D present frequently repeated tile-like shape (compare Sup. Figure 6-14)

that typically have to be separated into smaller pieces for data collection. Here, they are approximately 100x100 μm and 100x30 μm sizes.

Both X-ray data sets were collected at the D23-1 of the ESRF synchrotron centre. With 1° oscillations, 180 images were recorded using a PILATUS 6M-F detector with an exposure time of 0.1 (0.75) s per image and a crystal-to-detector distance of 501.56 (501.72) mm for PBGD1 P3' (PBGD2 P1). The incident beam had a wavelength of 0.972422 Å. The diffraction data obtained from the crystal were processed to a resolution of 2.74 (2.80) Å, revealing that both crystals belong to the monoclinic space group $P2_1$, with unit-cell parameters a = 68.80 Å, b = 81.09 Å, c = 79.39 Å, β = 93.299° (a = 70.67 Å, b = 81.11 Å, c = 76.58 Å, β = 95.270°); for details, see Table 4-4. By using the method of Matthews[166,167], it was estimated that both crystals contained a single PBGD dimer per crystallographic asymmetric unit with a solvent content of 47%. Analysis by molecular-replacement with the *H. sapiens* PBGD (PDB ID: 3ECR) as the search model was successful and the structures are currently undergoing the refinement process.

	Protein sample	**PBGD1 P3'**	**PBGD2 P1**
Beamline param.	Beamline	ID23-1 (ESRF)	ID23-1 (ESRF)
	Detector	Pilatus 6M-F	Pilatus 6M-F
Experiment param.	Wavelenght [Å]	0.972422	0.972422
	Detector distance [mm]	501.56	501.72
	No. of images, Oscilation range [°]	180, 1	180, 1
	Exsposure time (total) [s]	0.10 (18)	0.75 (135)
Data collection statistics	Space group	$P2_1$	$P2_1$
	Unit cell parameters	a= 68.80 b=81.09 c= 79.39 β= 93.299	a= 70.67 b= 81.11 c= 76.58 β= 95.270
	Resolution range [Å]	50-2.74 (2.91-2.74)	50-2.8 (2.97-2.8)
	Unique reflections	22646 (3593)	20705 (3001)
	Multiplicity	3.4 (3.3)	3.3 (3.3)
	Completness [%]	98.7 (97.7)	97.1 (88.2)
	R_{merge}	21.1 (149.5)	24.7 (120.0)
	R_{meas} [%]	25.1 (178.1)	29.3 (142.4)
	I/σ [I]	5.39 (0.94)	6.39 (1.80)
	CC(1/2)	98.6 (46.0)	97.5 (60.9)
	Molecule per AU	2	2
	Solvent content [%]	47	47
Data refinement statistics	No. Pf reflections (total/test)	In process	In process
	Rwork/Rfeww [%		
	No. Of protein atoms		
	Wilson/average B-factor [Å²]		
	RMS Δ bond/angle [Å, °]		
	Ramachandran statistics: Favoured/allowed/autliers [%]		

( ) last resolution shell

Table 4-4. Data collection parameters and processing statistics for PBGD1 P3' and PBGD2 P1.

Although R-factor values for the last shell are very high, we accept the resolution reaching 2.74 Å in case of PBGD1 P3' and 2.80 Å in case of PBGD2 P1 as all the other parameters are within the generally accepted standards, including R-factors of ~5% in the low resolution range (data not included into the Table 4-4). R_{merge} and R_{meas} are respectively 54.1 (50.3) % and 64.2 (60.3) for 3.11 (3.18) Å however, this resolution seems to be poor in the interaction exploring and active centre inspection point of view. Preliminary refined structures are shown in the Figure 4-32.; focus on the active centre.

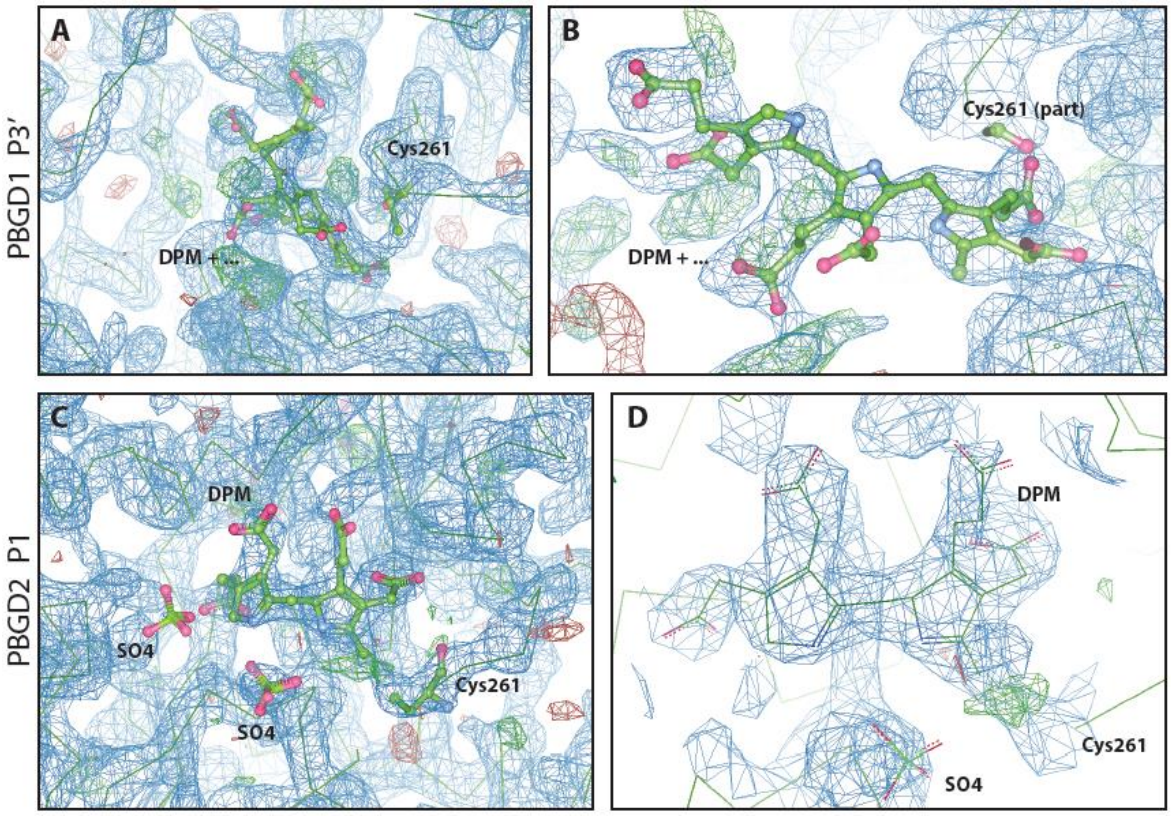


Figure 4-32. Active site of human PBGD; preliminary refinement. Cofactor, Cys261 residue and sulphate ions are represented as ball & sticks. (A-B) PBGD1 P3', (C-D) PBGD2 P1.

With first look-over to the refined models it becomes clear that active centre occupation differ between PBGD1 P3' (A) and PBGD2 P1 (C). In case of PBGD1 P3', DPM cofactor goes beyond it, elongated by yet undefined amount of PBG rings. In case of PBGD2 P1, DPM cofactor is held inside the active centre which entrance is blocked by one (or more sulphate ions). Both models differ also with the quality of density maps for the substrates: PBGD1 P3' (B) is difficult to model as a result of – most probably – flexible multi-pyrrole chain. On the other hand, cofactor was easy to modelled into the active centre of PBGD2 P1 (D), where the map adopted clearly its shape.

Contrary to what was expected but in agree to e.g. SAXS data analysis – there are neither significant domain movements nor re-arrangements of the loop covering the active centre. For broader summary, the data sets require further processing at the same time being the most promising from all the prior attempts to explain how growing tetrapyrrole chain is accommodated within the active site of PBGD enzymes.

4.2.3 UROIIIS destabilization mechanism

This section is directly based on the ben Bdira et al. article, “Tuning intracellular homeostasis of human uroporphyrinogen III synthase by enzyme engineering at a single hotspot of congenital erythropoietic porphyria”, where modelling issue was covered by me.

As early described, congenital erythropoietic porphyria (CEP) is a rare disease that results from a deficient activity in the uroporphyrinogen III synthase enzyme (UROIIIS) and it is ultimately produced by deleterious mutations in the *uroS* gene; more details in the section 1.4.4.
At the molecular level, over 30 mutations located in coding regions of UROIIIS have been identified from CEP patients[58]. Most of them correspond to missense mutations that reduce the *in vitro* catalytic activity, in most cases by lowering the enzyme's thermodynamic stability.[27] When analysed *in vivo*, the low stability found for the mutated uroporphyrinogen III synthase ultimately limited the enzyme's homeostasis: experiments in human cell lines [149] and with a murine model of the disease [150] have shown that the protein is fast eliminated partially via the proteasome. We have studied the structural basis for the enzyme destabilization at 73 hotspot position – the centre of attention was one of the most aggressive mutations, Cys73Arg .
Although Cys73 is not essential for catalysis [149], it is evolutionarily conserved: analysis of all the sequences available in the PUBMED gene repository shows that the vast majority of living organisms encode for a C (61%), S (22%) or A (14%) at this position. A strong correlation was found between the side chain volume at this position and the lifetime of the folded conformation. Mutations at position 73 modulate the inter-domain closure, ultimately affecting overall stability. Molecular modelling was used to rationalize these results, showing that the mutation site is coupled to the hinge region separating the two domains. By incorporating residues prone to interact with mutated amino 73 and to stabilize the hinge region, the catalytic activity was fully restored and a moderate increase in the kinetic stability of the enzyme was also observed.
Given the onerousness of the disease and the ineffectiveness of current therapies, molecular chaperons are seen as a new hope for CEP patients. They were reported to bind to the enzyme, stabilize its folded (active) conformation and thus up regulate protein's homeostasis [168]. In practice, this strategy is challenging since the interaction between the molecules has to be energetically productive (i. e. stabilizing the enzyme). Therefore, in order to provide the most effective outcome in chaperone-based therapies it is very important to understand the structural determinants of the destabilization upon mutation.

4.2.3.1 Residual catalytic activity in UROIIIS_Cys73Arg

It was recently shown that UROIIIS_Cys73Arg is kinetically unstable, with the substitution accelerating the protein’s unfolding ≈ 380 times when compared with wild-type (WT) UROIIIS [45]. Figure 4-33 shows catalytic activity assays testing WT and UROIIIS_Cys73Arg, and the derived kinetic constants are listed in Table 4-5.

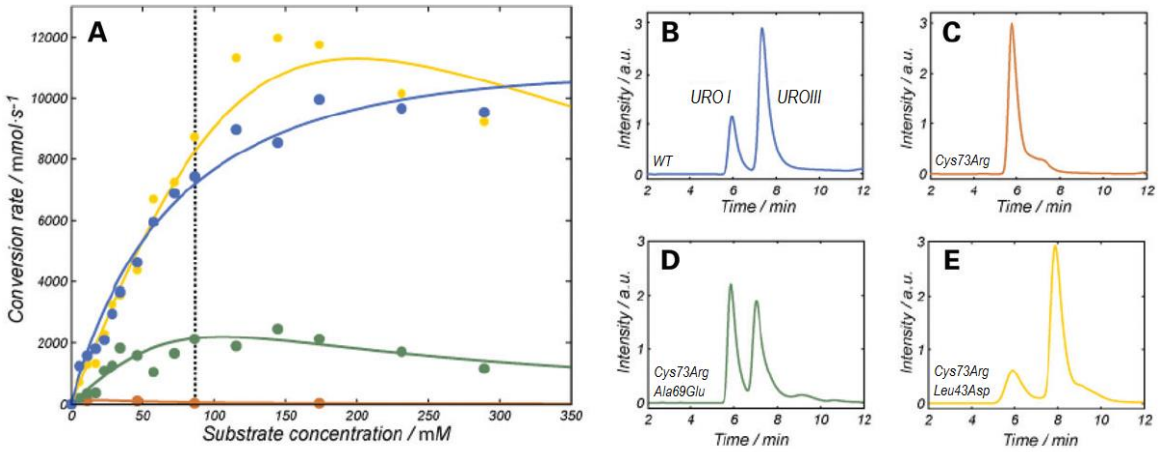


Figure 4-33. Catalytic activity for the engineered variants of UROIIIS. Left (A) Conversion rate as a function of the porphobilinogen concentration for WT (blue), UROIIIS_Cys73Arg (brown), Cys73Arg/Ala69Glu (green) and Cys73Arg/Leu43Asp (yellow). Solid lines correspond to the best fitting to a Michaelis–Menten equation (WT) or to a Michaelis–Menten with excess of substrate inhibition equation (mutants). The Michaelis constant (K_M), the catalytic rate (k_{cat}) and the inhibition constant (K_i) obtained from the fitting are listed in Table 4-5. Right (B–E) Chromatograms showing the HPLC separation between the enzymatic product (uroIII) and the by-product (uroI) from an enzymatic reaction mixture (obtained with 100 mM of substrate). Each plot has been obtained from a single protein variant, according to the left panel colour code.

	K_M (μM)	k_{cat} (s^{-1})	k_{cat}/K_M (M^{-1} s^{-1})	K_i (μM)
WT[a]	60 ± 20	890 ± 50	1.5×10^7	-
Cys73Arg[b]	> 10 000	10 ± 9	$< 10^3$	n.a.
Cys73Arg/Ala69Glu[b]	50 ± 20	170 ± 70	3.7×10^6	120
Cys73Arg/Leu43D[b]	70 ± 20	1200 ± 90	1.6×10^7	1500

Table 4-5. Enzymatic parameters for WT-UROIIIS and the mutants under consideration. ([a]) Values obtained from the fitting to a Michaelis–Menten model, ([b]) Values obtained from the fitting to a Michaelis–Menten model with inhibition by excess of substrate.

The enzymatic assay is able to determine the amount of enzymatic product (URO III) and the concentration of the by-product (URO I), which is spontaneously generated and ultimately responsible for most of the symptoms observed in CEP patients [169]. UROIIIS_WT follows a canonical Michaelis–Menten mechanism (Figure 4-33A, blue) and the catalytic efficiency is high enough to prevent the appearance of URO I (Figure 4-33B). Freshly purified UROIIIS_Cys73Arg shows residual activity that is best modelled when accounting for inhibition due to an excess of substrate (Figure 4-33A, brown), and has a catalytic efficiency 15000-fold weaker than WT-UROIIIS (Table 4-5) due to deterioration of the folded conformation during the enzymatic assay. Such low activity results in the generation of large amounts of URO I (Figure 4-33), consistent with the pathogenic character of this mutation.

4.2.3.2 A structural model for the hotspot UROIIIS destabilization

The crystal structure of UROIIIS_WT revealed that the protein is composed of two domains, connected by a hinge region[170]. In the absence of high resolution structures for the mutant proteins and the impossibility of obtaining them because of their modest stability, molecular

dynamics (MD) simulations in explicit solvent were used to generate realistic structural models that account for alteration in Cys73 (3.1.3 section). The time trajectories for WT and UROIIIS_Cys73Arg obtained from the molecular simulations were analysed to characterize the effect of the modifications on the folded conformation and compared them with the starting structure.

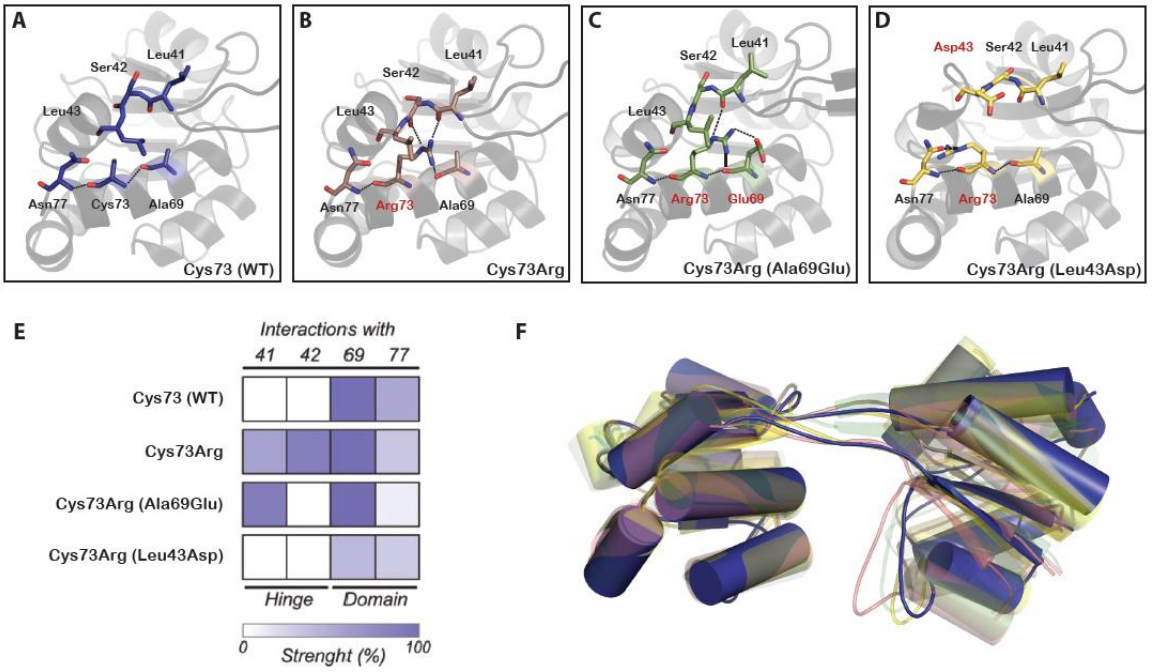


Figure 4-34. A model for the kinetic destabilization observed in Cys73R. (A–D) Representative structures for all the proteins under consideration obtained from the molecular dynamics simulations. Dashed lines underline the inter-residue interactions detected in the simulations. (E) Comparison of the strength of the detected interactions among the UROIIIS variants. The strength of the interaction is colour coded according to the legend and relative to the maximum strength found ever. (F) Overlay of the different structures aligned by the C-domain to highlight differences in the inter-domain orientation. Colour code: WT (blue), Cys73Arg (brown), Ala69Glu/Cys73Arg (green) and Leu43Asp/Cys73Arg (yellow).

UROIIIS_WT shows a high degree of conservation of the native structure during the simulation's trajectory, only deviating from the 1JR2 structure [43] in the orientation between the two domains (Figure 4-34B). This result is in full agreement with a previous crystallographic study showing that UROIIIS adopts multiple inter-domain orientations with a preference for extended structures [171]. In UROIIIS_WT, Cys73 interacts (intra-domain) with the nearby residues Asn77 and Ala69 (Figure 4-34A and E), but remains totally uncoupled from residues of the hinge region. In UROIIIS_Cys73Arg (Figure 4-34B), Arg73 is repositioned in a conformation that avoids steric clash with Leu43, providing a structural model to explain the experimental correlation observed between the reduced stability and the side chain volume at position 73. Moreover, the introduced arginine creates two new interactions with residues from the hinge region (Leu41 and Ser42), while the interaction between Arg73 and Asn77 is weakened (Figure 4-34E). Cys73Arg-UROIIIS shows a twisting motion of the two domains when compared with UROIIIS_WT (of ~32° on average, Figure 4-34F), probably as a result of the interaction between Arg73 and Leu41-Ser42. We hypothesize that the introduction of bulky residues at position 73 (like arginine) forces the side chain to form noncovalent interactions with the hinge region, twisting the domains and ultimately destabilizing the protein.

4.2.3.3 Engineering the hotspot stability site

We have engineered the residues surrounding Arg73 and evaluated their capability to rescue the catalytic properties of the enzyme. Considering that the Cys73Arg substitution introduces a positive charge, our engineering strategy aimed to alter the conformation of Arg73 through the addition of negatively charged residues. Molecular dynamics simulations were used to model the interaction pattern of Arg73 and the new pocket configuration. First, we considered the introduction of a glutamic acid at position 69 (UROIIIS_Cys73Arg/Ala69Glu) that, according to MD simulations, should shift the orientation of Arg73 by ~30° (in chi3, Figure 4-34C), abrogating the interaction with Leu41 (Figure 4-34E) and reducing the inter-domain twist by ~11° (Figure 4-34F). As shown in Figure 4-33A (green), UROIIIS_Cys73Arg/Ala69Glu is able to partially restore the catalytic activity of UROIIIS_Cys73Arg. Specifically, the Ala69Glu mutation enhanced the catalytic activity by a factor of 16, showing little effect on the formation of the Michaelis complex (Table 4-5), as expected, since it is far from the substrate-binding site located in the cleft between the two domains [172]. Next, we considered the introduction of a negative charge at position 43 (UROIIIS_ Cys73Arg/Leu43Asp). This mutant results in electro neutralization of the pocket but, according to the structural model, no interactions are created between Asp43 and Arg73 (Figure 4-34D). Instead, the pocket becomes sterically relaxed, largely reconstituting the interaction pattern of UROIIIS_WT. When assayed, UROIIIS_ Cys73Arg/Leu43Asp showed a larger increase in kcat than UROIIIS_Cys73Arg/Ala69Glu (Figure 4-33A, yellow). Remarkably, this construct fully restores the catalytic activity and its catalytic efficiency is comparable (within experimental error) with that of UROIIIS_WT (Table 4-5).

As the drop in the catalytic activity over time is always produced by a loss in the secondary structure, it can be followed by circular dichroism experiments. The time dependence of the catalytic activity at physiological temperature for all the enzyme variants under consideration proved that in the timeframe of the experiment (< 25 h) wild type's activity remained largely unaffected while UROIIIS_Cys73Arg catalytic activity was lost completely in less than 1h. For UROIIIS_Cys73Arg/Leu43Asp, a significant decay over time was observed in the catalytic power, lowering to a 50% of activity in 12h. UROIIIS_Cys73Arg/Ala69Glu showed a slower decay over time reaching half-life activity in about 16h.[151]

C-terminal Myc-tagged-versions of the four proteins were also expressed in the human M1 cell line and their expression levels were analysed by Western blotting. A fraction of cells incubated with proteasome inhibitor MG132 were further lysed to test enzymatic activity of UROIIIS. The Cys73Arg and Cys73Arg/Leu43Asp mutants retain about 15% of the WT activity, while UROIIIS_Cys73Arg/Ala69Glu enhances the Cys73Arg activity until 30% (normalized to the WT). These results suggest that the expression of the double mutant UROIIIS_Cys73Arg/Ala69Glu protein in conditions that favour their accumulation was able to increase significantly the intracellular activity of UROIIIS_Cys73Arg.[151]

After systematic analysis of the protein variants at position 73 it is now clear that the loss of stability found in the pathogenic Cys73Arg is produced by the steric clashing between the (bulky) arginine side chain and nearby atoms. The excellent correlation found for the side chain volume at

this position and the agreement with the phylogenetic conservation underlines the relevance of this site in maintaining the protein integrity and its role in protein intracellular homeostasis regulation. *In silico* calculations provide a plausible model for the rearrangement of the interaction network upon mutation and suggest a coupling between position 73 and the hinge residues. In bi-domain proteins, the composition and structure of the hinge is often pivotal for the stability and functionality of the protein [173]. This hypothesis has been subsequently tested with rescue mutations: second-site mutations designed to abrogate such coupling and to restitute the interaction network observed for wild type UROIIIS (Figure 4-34) succeeded in rescuing wild type' folding protein properties and the activity defects caused by the deleterious mutation *in vitro*.

The double mutants that successfully rescue the enzyme activity also showed a partial recovery of the kinetic stability, *in vitro*. Actually, for this enzyme, an increase in the protein's (kinetic) stability is almost always accompanied by an increase in the catalytic efficiency, consistent with the fact that the enzyme lacks a truly active site and a large number of residues are involved in the catalytic mechanism [170,172]. Interestingly, one of these double mutants is able to significantly enhance the enzymatic activity also in eukaryotic cells supporting that a target therapy through molecular chaperones that emulates the effect of the Ala69Glu on the UROIIIS_Cys73Arg mutant protein could be beneficial for CEP patients. However, in contrast to the protein expression found in bacterial systems, the expression in a mammalian cellular environment of the double mutants (Cys73Arg/Leu43Asp and Cys73Arg/Ala69Glu) are not sufficient to slow down the intracellular degradation caused by the single Cys73Arg mutation. This could be due to the fact that mammalian cells display more complex mechanisms (e.g. endoplasmic-reticulum-associated protein degradation (ERAD)) for protein quality control than bacteria systems. These findings highlight the need for further investigations to unravel the pathways involved in the UROIIIS intracellular degradation in order to increase the success rate of the molecular chaperone-based therapies.

PERSPECTIVES

In the future, additional structures of enzyme-substrate and enzyme-inhibitor complexes followed by robust simulations will certainly provide even more insight into the mechanisms of enzyme catalysed heme biosynthesis.

CONCLUDING REMARKS

5

5.1 Connexins

- In the present contribution I have used computational modelling to generate, for the first time, a large set of structural models for homotypic, heterotypic and heteromeric GJs from Cx26, Cx32 and Cx47. The agreement with *ad hoc* experimental data validates such models. This plethora of structures allowed a thorough comparison to unravel the structural determinants for GJ formation.

- It was found that the computational docking energy can be considered as a reporter for the docking strength and it reflects the energetic complementarity between the two hemichannels in the junction. The agreement found between this parameter and the experimental yields for GJ formation (Figure 4-7, Figure 4-8) reveals that this is a key factor for the junction formation, also explaining important aspects of the connexin promiscuity. Moreover, such agreement validates the novel methods that have been used.

- My studies confirm the relevance of the extracellular loops in the channel formation and the crucial role of residue Asn176 of Cx26 for hemichannel recognition (and their corresponding ones in the other Cx isoforms) and residue Asn54 of Cx26 for channel stability. They also reveal that hemichannel composition is only partially responsible for the channel formation, while the relative disposition of connexins in the hemichannels ultimately define the hemichannel compatibility.

- Analysis of the hemichannel interactions, based on rigid docking procedure run for almost 2000 complexes, provides new and valuable features about gap junction specificity and the deleterious effect of the pathogenic mutants. It also suggests hydrogen bond formation at the docking interface more diverse as previously reported[24,111].

- Supported with functional studies, analysis of MD trajectories for WT Cx47 hemichannel and its altered versions proved the importance of N-terminal domain and first transmembrane domain in the proper functioning of hemichannels. Simulations, carried out for the first time for Cx47 protein, indicated residues crucial for the funnel pore organization. Also, reported the modifications of the pore architecture (thus, hemichannel permeability) resulting from the disruption of the interaction network.

5.2 Heme pathway enzymes

- Molecular docking analysis resulted in a list of compounds mimicking the substrate and exhibiting inhibition potential. It also allowed to spot and characterize two target site of inhibitory action ("Pocket 1" and "Pocket 2").

- Enzymatic assays with the selected compound indicated competitive inhibition and thus, confirms the relevance of the molecular docking assumptions and analysis. Moreover, these results indicate that PBGD can be effectively inhibited by small organic molecules; tested compounds constitute a new line of inhibitors from which 2,5-dimethyl-1H-pyrrole demonstrating the highest level of inhibition in the relation to its concentration in the reaction solution has been established as a reference for further trials.

- Modified PBGD preparation protocol allow to separate and monitor stable reaction intermediates apt for the structural studies which results prove the dynamic character of the enzyme and results in the first diffraction of intermediates.

- The investigation of second-site mutations that suppress aberrant alleles (carried out on the example of UROIIIS protein) is a useful strategy for the unbiased identification of residues involved in quaternary assembly and of structural regions that might be manipulated to restore wild-type function.

- UROIIIS studies unravelled the delicate interplay between Cys73 and the kinetic stability of the protein and proved that non-covalent electrostatic interactions at this site are able to partially restore the enzyme stability and activity.

- Both biochemical and structural characterization of the heme biosynthesis enzymes help to understand the molecular basis of enzyme mechanisms and the deleterious effect of porphyria-causing mutations, providing the basic tools to develop new strategies for the treatment of these severe metabolic disorders as well as may serve to establish new photodynamic cancer therapy (taking into account light sensitivity of porphyrins) and design new antibiotics or herbicides (focusing on enzymes involved in heme biosynthesis exclusively in plants or bacteria).

APPENDICES

6

Abbreviations

AIR – acute intermittent porphyria (AIP)

CEP – congenital erythropoietic porphyria (CEP)

Cx(s) – connexin(s)

cryoER – cryo-electron microscopy

ALA – D-aminolevulinic acidDPM – dipyrromethane cofactor

ER – endoplasmic reticulum

EDTA – ethylenediaminetetraacetic acid

FFT – fast Fourier transform

FF – force field

GJ(s) – gap junctions

LGA – Lamarckian Genetic Algorithm

LwB – Lineweaver-Burk plot

MD – molecular dynamics

MM – molecular mechanics

MTS – multiple time stepping

NMR – nuclear magnetic resonance spectroscopy

PBG – porphobilinogen

PBGD – Porphobilinogen Deaminase

HMP – pre-uroporphyrinogen (1-hydroxymethylbilane

QM – quantum mechanics

URO I, Uro'gen I – uroporphyrinogen I

URO III, Uro'gen 3 – uroporphyrinogen III

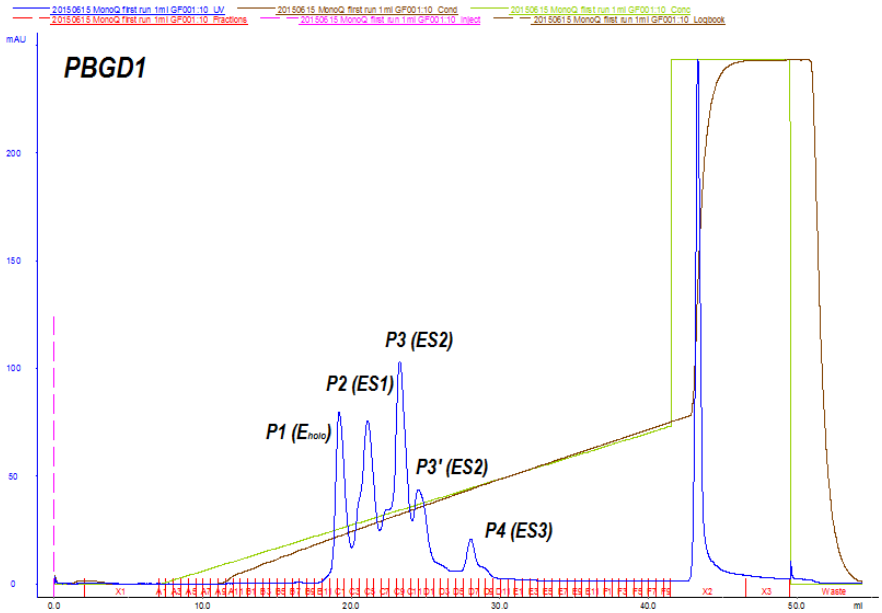
UROIIIS – Uroporphyrinogen III Synthase

NVT – volume-constant temperature

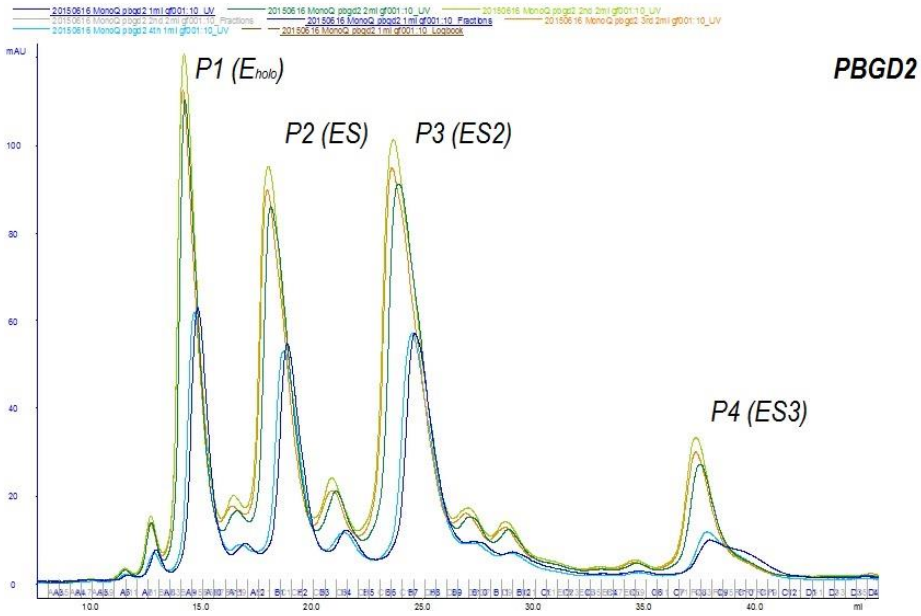
WT, wt – wild-type

Suplementary Figures

Anion exchange chromatograms

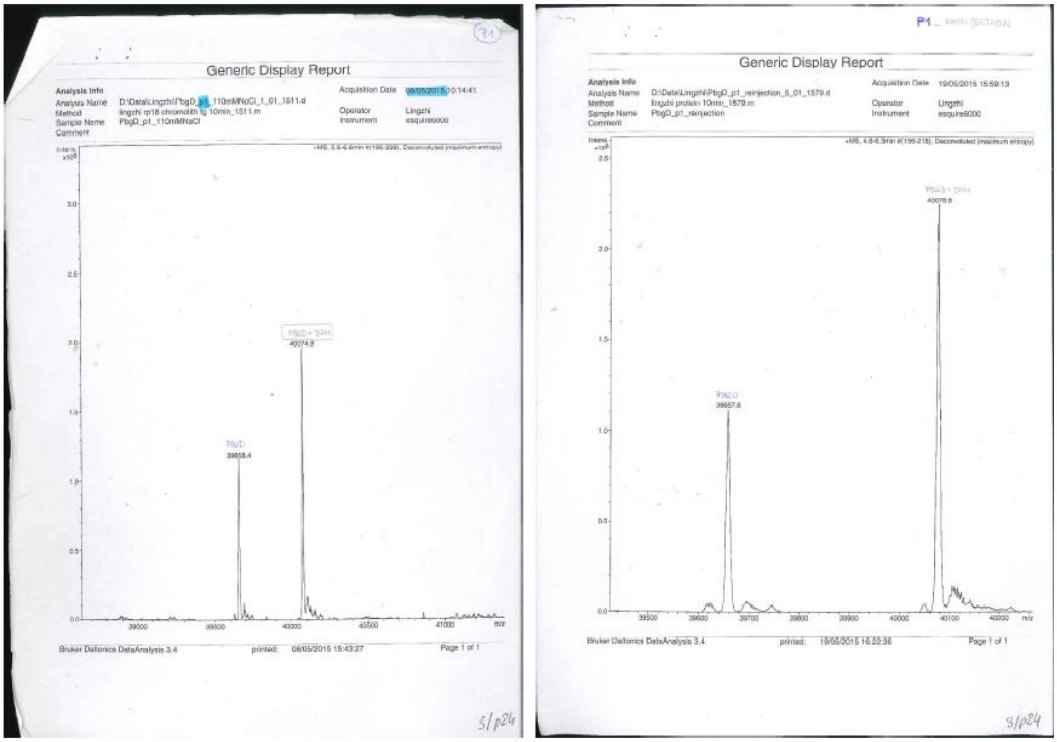


Sup. Figure 6-1. Anion exchange chromatography of wt PBGD1 in MonoQ 5/50 GL with samples corresponding to the peaks of the successive reaction stages, absorbance in arbitrary units (vertical axis), horizontal axis shows the elution volume and the elution fraction.

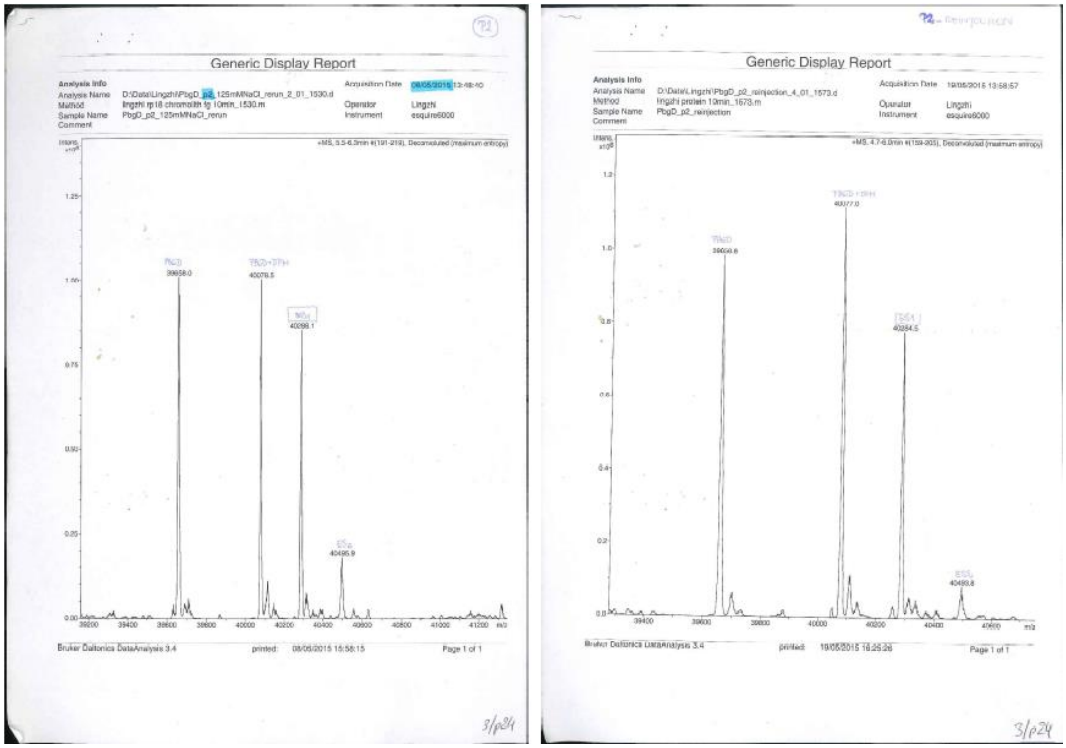


Sup. Figure 6-2. Anion exchange chromatography of PBGD2 in MonoQ 5/50 GL with samples corresponding to the peaks of the successive reaction stages. To not overload the column, samples were divided into 2 ml (green) and 1 ml (blue) portions and purified in separated runs. Superimposition proves the reproducibility of results in terms of conditions – samples comes from exactly the same preparation and runs are carried in short time period.

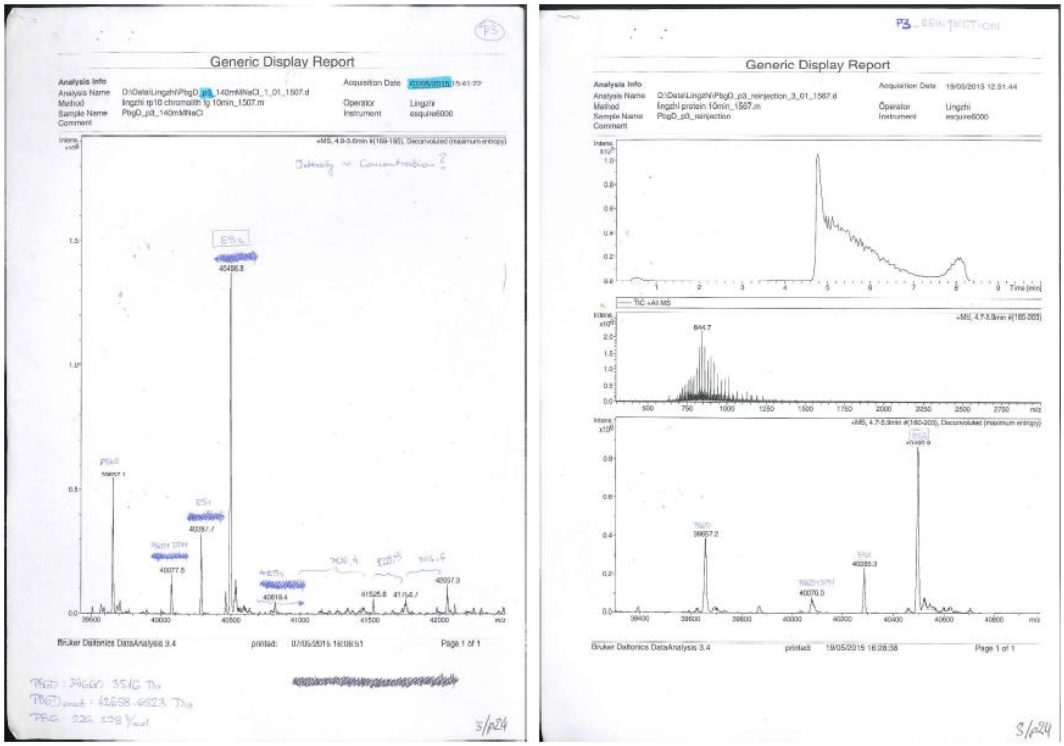
Mass-spec results



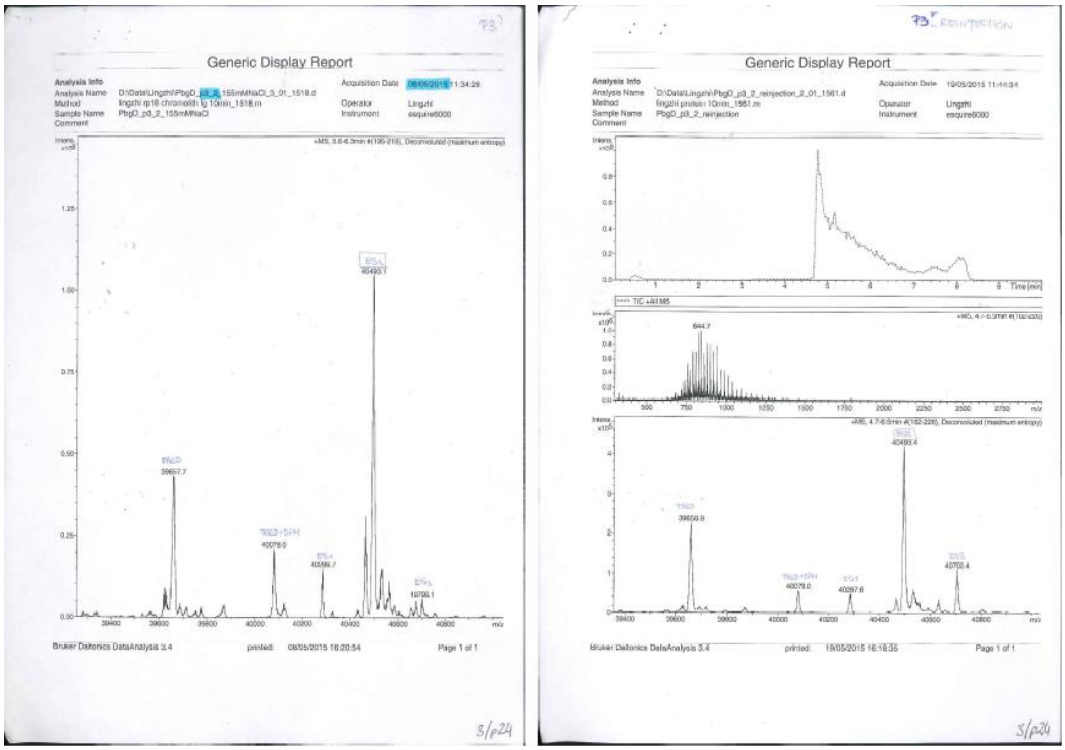
Sup. Figure 6-3. Mass-spec analysis of PBGD1 P1 sample composition for purified (left, full protocol procedure) and re-purified (right, Q column re-injection of "left" sample) protein.



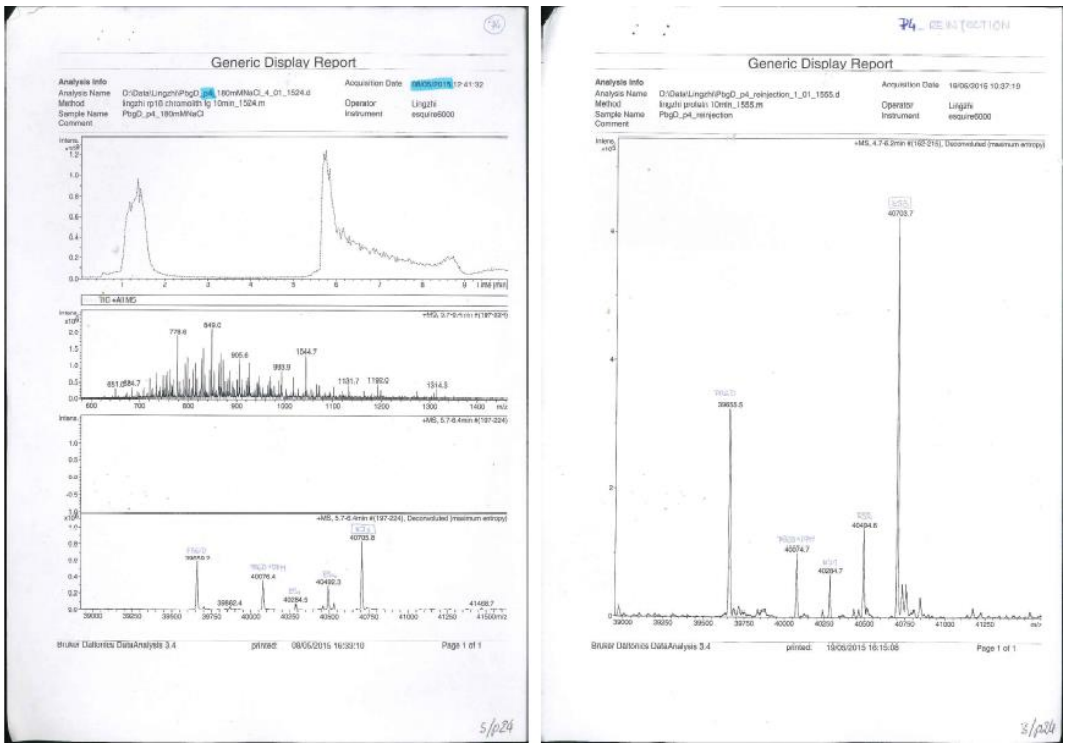
Sup. Figure 6-4. Mass-spec analysis of PBGD1 P2 sample composition for purified (left, full protocol procedure) and re-purified (right, Q column re-injection of "left" sample) protein.



Sup. Figure 6-5. Mass-spec analysis of PBGD1 P3 sample composition for purified (left, full protocol procedure) and re-purified (right, Q column re-injection of "left" sample) protein.

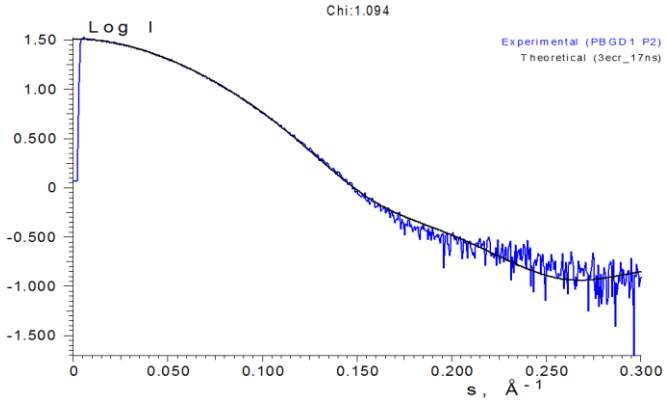


Sup. Figure 6-6. Mass-spec analysis of PBGD1 P3' sample composition for purified (left, full protocol procedure) and re-purified (right, Q column re-injection of "left" sample) protein.

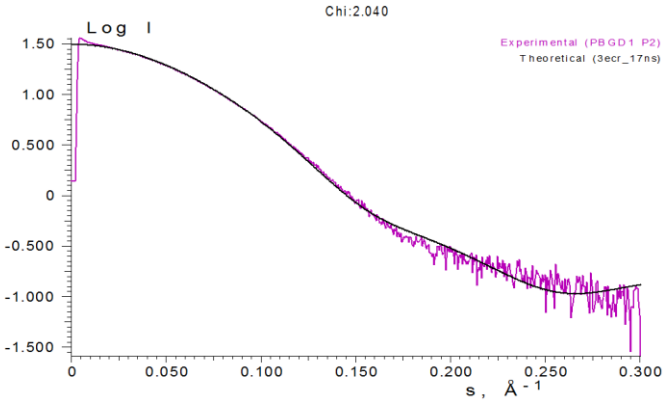


Sup. Figure 6-7. Mass-spec analysis of PBGD1 P4 sample composition for purified (left, full protocol procedure) and re-purified (right, Q column re-injection of "left" sample) protein.

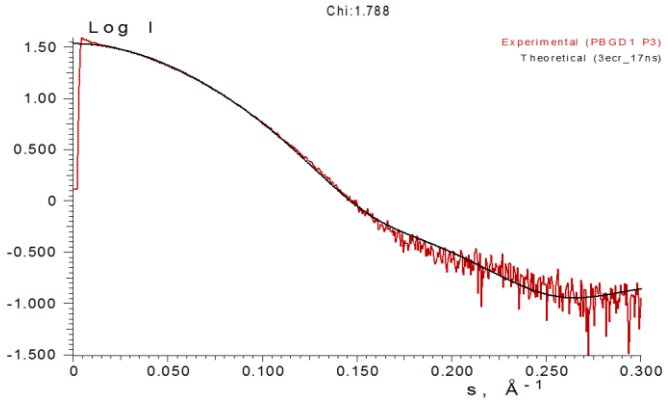
SAXS experimental spectra vs. simulated fits obtained from the model



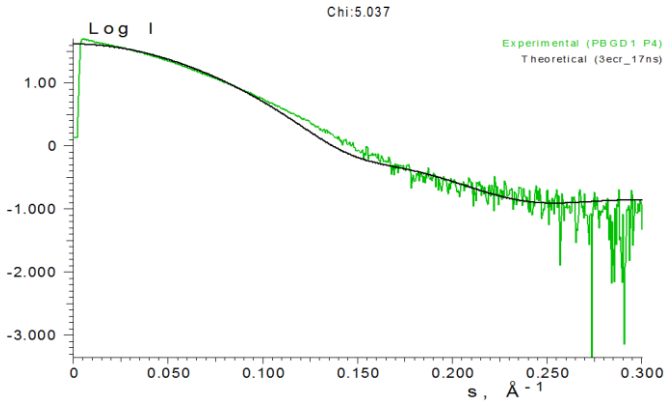
Sup. Figure 6-8. SAXS of the PBGD intermediates: PBGD1 P1, the experimental spectrum (blue) is shown with a simulated fit obtained from the relaxed crystal structure of with PDB ID: 3ECR (black). Chi value shown at the top.



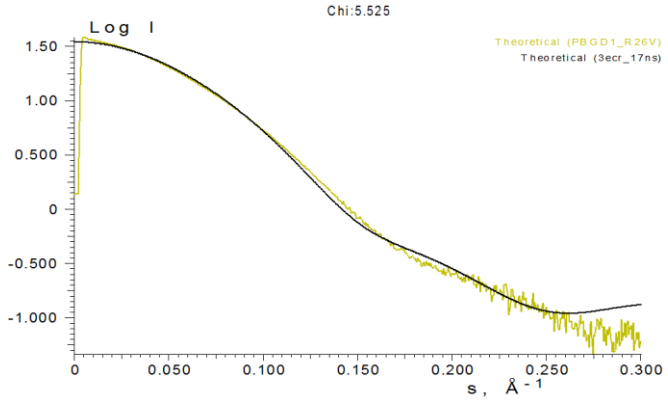
Sup. Figure 6-9. SAXS of the PBGD intermediates: PBGD1 P2, the experimental spectrum (magenta) is shown with a simulated fit obtained from the relaxed crystal structure of with PDB ID: 3ECR (black). Chi value shown at the top.



Sup. Figure 6-10. SAXS of the PBGD intermediates: PBGD1 P3, the experimental spectrum (red) is shown with a simulated fit obtained from the relaxed crystal structure of with PDB ID: 3ECR (black). Chi value shown at the top.

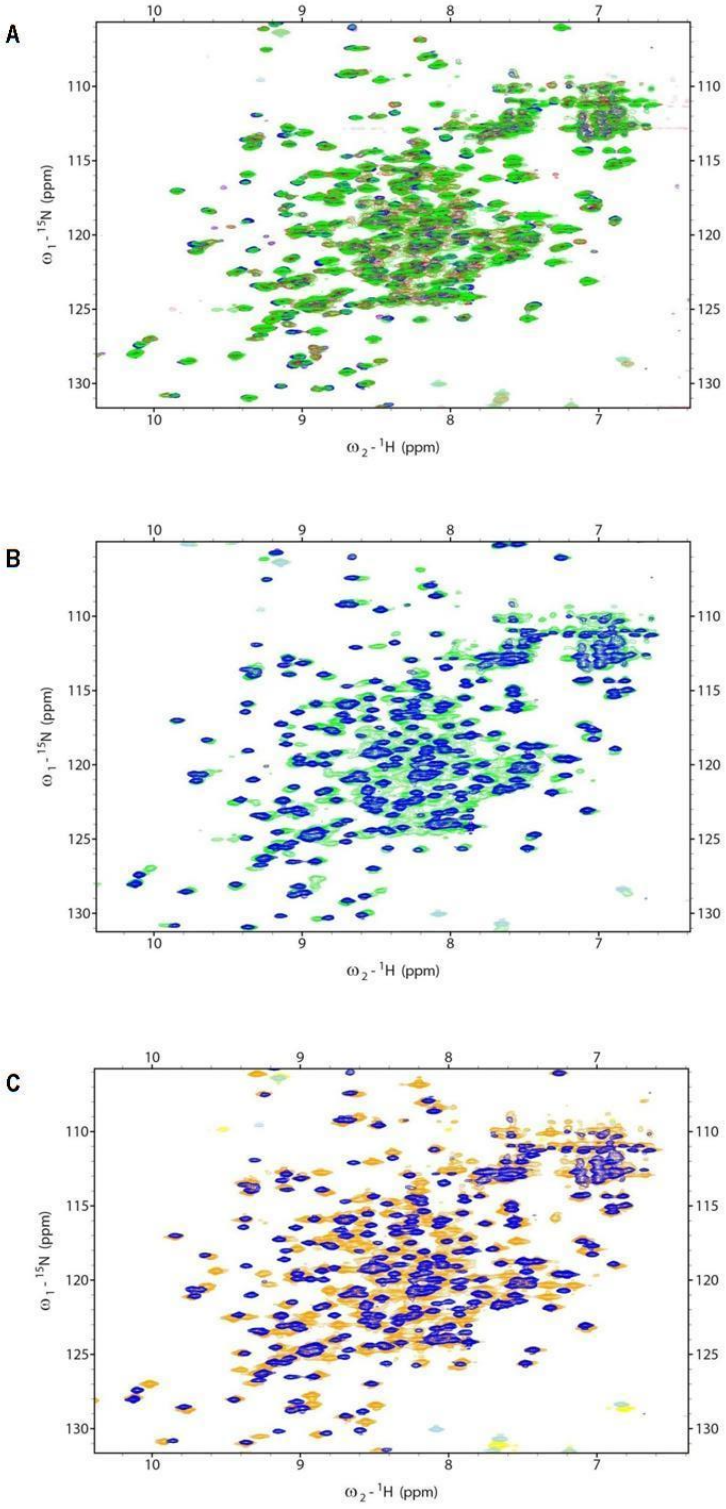


Sup. Figure 6-11. SAXS of the PBGD intermediates: PBGD1 P4, the experimental spectrum (green) is shown with a simulated fit obtained from the relaxed crystal structure of with PDB ID: 3ECR (black). Chi value shown at the top.



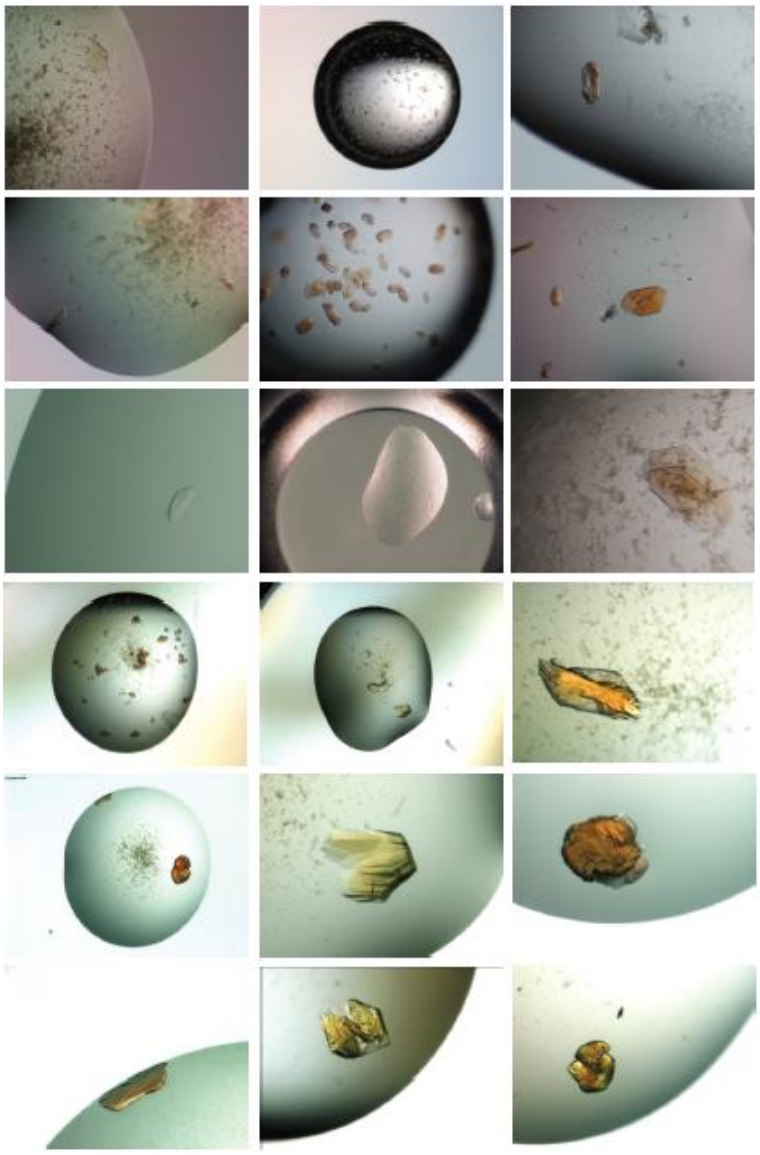
Sup. Figure 6-12. SAXS of the PBGD mutant: PBGD1_R26V, the experimental spectrum (yellow) is shown with a simulated fit obtained from the relaxed crystal structure of with PDB ID: 3ECR (black). Chi value shown at the top.

$^{1}H,^{15}N$-HSQC NMR Spectra



Sup. Figure 6-13. $^{1}H,^{15}N$-HSQC spectra for PBGD reaction intermediates and R26V mutant. Superimposition of the spectra for: (A) PBGD1 P1-P4, (B) PBGD1 P1 and P4, (C) PBGD1 P1 and PBGD1_R26V. Colours: (dark blue) P1, (magenta) P2, (red) P3, (green) P4, (orange) R26V mutant.

X-ray Crystallography data



Sup. Figure 6-14 PBGD crystal examples.

Protein sample	**PBGD1 (P3')**	**PBGD2 (P1)**
Protein concentration	estim. 5-10mg/ml	estim. 5-10mg/ml
Reservoir solution	0.1 M Na HEPES pH 7.2 (buf.), 0.9 M Na dihyhydrogen phosphate/ 0.9 M K dihydrogen phosphate (salt); Based on Molecular Dimension screen: JCSG-plus™ HT-96 screen, weel #C5	0.1 M sodium cacodylate pH 6.5 (buf.), 0.2 M sodium chloride (salt), 2 M ammonium sulphate (salt); Molecular Dimension screen: JCSG-plus™ HT-96 screen, weel #E2
Cryoprotectant	25% glycerol	25% glycerol
Plate	Linbro 24-well; put by hand	SBS 96-well flat bottom; put by Mosquito™ robot
Reservoir volume [µL]	500	200
Protein drop volume [µL]	1	0.2
Protein to solution ratio	1:1	1:1
Temperature [°C]	21 (in dark)	21 (in dark)
Set date	19/07/2015	16/06/2015
Fishing date	20/07/2015	20/07/2015
Data collection date	24/07/2015	24/07/2015

Sup. Table 6-1. PBGD crystallization conditions

List of Equations

List of Figures

List of Tables

References

1 What is structural biology?, Instruct Integratin Biology, http://www.structuralbiology.eu/ (2011)
2 Bergethon PR (2010) *The Physical Basis of Biochemistry: The Foundations of Molecular Biophysics* Springer Science & Business Media.
3 Nelson DL & Cox MM (2013) *Lehninger Principles of Biochemistry* W.H. Freeman.
4 Berg JM, Tymoczko JL & Stryer L (2002) *Biochemistry, Fifth Edition: International Version* Granite Hill Publishers.
5 Leach AR (2001) *Molecular Modelling: Principles and Applications* Pearson Education.
6 Campbell NA, Williamson B & Heyden RJ (2006) *Biology: Exploring Life* Pearson Prentice Hall.
7 Mancin F Chimica Supramolecolare. *Supramolecular Chemistry and Functional Nanoparticles*.
8 Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D & Darnell J (2000) *Molecular Cell Biology*, 4th ed. W. H. Freeman.
9 Bosshard HR, Marti DN & Jelesarov I (2004) Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. *J. Mol. Recognit.* **17**, 1–16.
10 Teaching & Learning Packages, University of Cambridge, http://www.doitpoms.ac.uk/
11 Pace CN, Shirley BA, McNutt M & Gajiwala K (1996) Forces contributing to the conformational stability of proteins. *FASEB J* **10**, 75–83.
12 Bissantz C, Kuhn B & Stahl M (2010) A Medicinal Chemist's Guide to Molecular Interactions. *J Med Chem* **53**, 5061–5084.
13 Jones S & Thornton JM (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A* **93**, 13–20.
14 Seitz H (2007) *Analytics of Protein-DNA Interactions* Springer Science & Business Media.
15 *Thermo Scientific Pierce Protein Interaction Technical Handbook* (2010) Thermo Fisher Scientific Inc.
16 Sanders CR Biomolecular Ligand-Receptor Binding Studies: Theory, Practice, and Analysis. .
17 Dunn MF (2001) Protein–Ligand Interactions: General Description. In *eLS* John Wiley & Sons, Ltd.
18 Alberts B, Johnson A, Lewis J, Raff M, Roberts K & Walter P (2002) *Molecular Biology of the Cell*, 4th ed. Garland Science.
19 Cell Adhesion and Cell Communication *Scitable - A Collaborative Learning Space for Science*.
20 Meşe G, Richard G & White TW (2007) Gap Junctions: Basic Structure and Function. *J Invest Dermatol* **127**, 2516–2524.
21 Segretain D & Falk MM (2004) Regulation of connexin biosynthesis, assembly, gap junction formation, and removal. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1662**, 3–21.
22 Verheule S & Kaese S (2013) Connexin diversity in the heart: insights from transgenic mouse models. *Front Pharmacol* **4**.
23 Harris AL (2001) Emerging issues of connexin channels: biophysics fills the gap. *Q. Rev. Biophys.* **34**, 325–472.
24 Maeda S, Nakagawa S, Suga M, Yamashita E, Oshima A, Fujiyoshi Y & Tsukihara T (2009) Structure of the connexin 26 gap junction channel at 3.5 Å resolution. *Nature* **458**, 597–602.
25 Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD & Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1–9.
26 Lai-Cheong JE, Arita K & McGrath JA (2007) Genetic Diseases of Junctions. *J Invest Dermatol* **127**, 2713–2725.
27 Layer G, Reichelt J, Jahn D & Heinz DW (2010) Structure and function of enzymes in heme biosynthesis. *Protein Sci* **19**, 1137–1161.

28 Heinemann IU, Jahn M & Jahn D (2008) The biochemistry of heme biosynthesis. *Arch. Biochem. Biophys.* **474**, 238–251.
29 Ajioka RS, Phillips JD & Kushner JP (2006) Biosynthesis of heme in mammals. *Biochim. Biophys. Acta* **1763**, 723–736.
30 Poblete-Gutiérrez P, Wiederholt T, Merk HF & Frank J (2006) The porphyrias: clinical presentation, diagnosis and treatment. *Eur J Dermatol* **16**, 230–240.
31 Porphyrins & Porphyria diagnosis, Sensitivity & specificity of tests for Porphyria *American Porphyria Foundation*.
32 Battersby AR (2000) Tetrapyrroles: the pigments of life. *Nat Prod Rep* **17**, 507–526.
33 Bogorad L (1958) The Enzymatic Synthesis of Porphyrins from Porphobilinogen I. UROPORPHYRIN I. *J. Biol. Chem.* **233**, 501–509.
34 Song G, Li Y, Cheng C, Zhao Y, Gao A, Zhang R, Joachimiak A, Shaw N & Liu Z-J (2009) Structural insight into acute intermittent porphyria. *FASEB J* **23**, 396–404.
35 Gill R, Kolstoe SE, Mohammed F, Al d-Bass A, Mosely JE, Sarwar M, Cooper JB, Wood SP & Shoolingin-Jordan PM (2009) Structure of human porphobilinogen deaminase at 2.8 Å: the molecular basis of acute intermittent porphyria. *Biochemical Journal* **420**, 17–25.
36 Jordan PM, Thomas SD & Warren MJ (1988) Purification, crystallization and properties of porphobilinogen deaminase from a recombinant strain of Escherichia coli K12. *Biochem J* **254**, 427–435.
37 Bung N, Pradhan M, Srinivasan H & Bulusu G (2014) Structural Insights into E. coli Porphobilinogen Deaminase during Synthesis and Exit of 1-Hydroxymethylbilane. *PLoS Comput Biol* **10**, e1003484.
38 Warren MJ & Scott AI (1990) Tetrapyrrole assembly and modification into the ligands of biologically functional cofactors. *Trends Biochem. Sci.* **15**, 486–491.
39 Grandchamp B, De Verneuil H, Beaumont C, Chretien S, Walter O & Nordmann Y (1987) Tissue-specific expression of porphobilinogen deaminase. Two isoenzymes from a single gene. *Eur. J. Biochem.* **162**, 105–110.
40 Louie GV, Brownlie PD, Lambert R, Cooper JB, Blundell TL, Wood SP, Warren MJ, Woodcock SC & Jordan PM (1992) Structure of porphobilinogen deaminase reveals a flexible multidomain polymerase with a single catalytic site. *Nature* **359**, 33–39.
41 Woodcock SC & Jordan PM (1994) Evidence for participation of aspartate-84 as a catalytic group at the active site of porphobilinogen deaminase obtained by site-directed mutagenesis of the hemC gene from Escherichia coli. *Biochemistry* **33**, 2688–2695.
42 Layer G, Reichelt J, Jahn D & Heinz DW (2010) Structure and function of enzymes in heme biosynthesis. *Protein Sci* **19**, 1137–1161.
43 Mathews MAA, Schubert HL, Whitby FG, Alexander KJ, Schadick K, Bergonia HA, Phillips JD & Hill CP (2001) Crystal structure of human uroporphyrinogen III synthase. *EMBO J* **20**, 5832–5839.
44 Schubert HL, Phillips JD, Heroux A & Hill CP (2008) Structure and mechanistic implications of a uroporphyrinogen III synthase-product complex. *Biochemistry* **47**, 8648–8655.
45 Fortian A, Castaño D, Ortega G, Laín A, Pons M & Millet O (2009) Uroporphyrinogen III synthase mutations related to congenital erythropoietic porphyria identify a key helix for protein stability. *Biochemistry* **48**, 454–461.
46 Kaur MD, Hazarika N, Saraswat N & Sood R (2015) Acute intermittent porphyria: Diagnostic dilemma and treatment options. *J Anaesthesiol Clin Pharmacol* **31**, 407–408.
47 Ramanujam V-MS & Anderson KE (2015) Porphyria Diagnostics-Part 1: A Brief Overview of the Porphyrias: Porphyria Diagnostics Overview. In *Current Protocols in Human Genetics* (Haines JL, Korf BR, Morton CC, Seidman CE, Seidman JG, & Smith DR, eds), pp. 17.20.1–17.20.26. John Wiley & Sons, Inc., Hoboken, NJ, USA.
48 Besur S, Hou W, Schmeltzer P & Bonkovsky H (2014) Clinically Important Features of Porphyrin and Heme Metabolism and the Porphyrias. *Metabolites* **4**, 977–1006.
49 Ged C, Moreau-Gaudry F, Richard E, Robert-Richard E & de Verneuil H (2009) Congenital erythropoietic porphyria: mutation update and correlations between genotype and phenotype. *Cell. Mol. Biol. (Noisy-le-grand)* **55**, 53–60.
50 Egan DN, Yang Z, Phillips J & Abkowitz JL (2015) Inducing iron deficiency improves erythropoiesis and photosensitivity in congenital erythropoietic porphyria. *Blood* **126**, 257–261.

51 Piomelli S, Poh-Fitzpatrick MB, Seaman C, Skolnick LM & Berdon WE (1986) Complete suppression of the symptoms of congenital erythropoietic porphyria by long-term treatment with high-level transfusions. *N. Engl. J. Med.* **314**, 1029–1031.
52 Katugampola RP, Anstey AV, Finlay AY, Whatley S, Woolf J, Mason N, Deybach JC, Puy H, Ged C, de Verneuil H, Hanneken S, Minder E, Schneider-Yin X & Badminton MN (2012) A management algorithm for congenital erythropoietic porphyria derived from a study of 29 cases. *Br. J. Dermatol.* **167**, 888–900.
53 Tishler PV & Winston SH (1990) Rapid improvement in the chemical pathology of congenital erythropoietic porphyria with treatment with superactivated charcoal. *Methods Find Exp Clin Pharmacol* **12**, 645–648.
54 Hift RJ, Meissner PN & Kirsch RE (1993) The effect of oral activated charcoal on the course of congenital erythropoietic porphyria. *Br. J. Dermatol.* **129**, 14–17.
55 Minder EI, Schneider-Yin X & Möll F (1994) Lack of effect of oral charcoal in congenital erythropoietic porphyria. *N. Engl. J. Med.* **330**, 1092–1094.
56 Gorchein A, Guo R, Lim CK, Raimundo A, Pullon HW & Bellingham AJ (1998) Porphyrins in urine, plasma, erythrocytes, bile and faeces in a case of congenital erythropoietic porphyria (Gunther's disease) treated with blood transfusion and iron chelation: lack of benefit from oral charcoal. *Biomed. Chromatogr.* **12**, 350–356.
57 Dupuis-Girod S, Akkari V, Ged C, Galambrun C, Kebaïli K, Deybach J-C, Claudy A, Geburher L, Philippe N, de Verneuil H & Bertrand Y (2005) Successful match-unrelated donor bone marrow transplantation for congenital erythropoietic porphyria (Günther disease). *Eur. J. Pediatr.* **164**, 104–107.
58 Ged C, Moreau-Gaudry F, Richard E, Robert-Richard E & de Verneuil H (2009) Congenital erythropoietic porphyria: mutation update and correlations between genotype and phenotype. *Cell. Mol. Biol. (Noisy-le-grand)* **55**, 53–60.
59 Richard E, Robert-Richard E, Ged C, Moreau-Gaudry F & de Verneuil H (2008) Erythropoietic porphyrias: animal models and update in gene-based therapies. *Curr Gene Ther* **8**, 176–186.
60 Use of inhibitors of porphobilinogen deaminase in the treatment of congenital erythropoietic porphyria
61 Freesemann AG, Gross U, Bensidhoum M, de Verneuil H & Doss MO (1998) Immunological, enzymatic and biochemical studies of uroporphyrinogen III-synthase deficiency in 20 patients with congenital erythropoietic porphyria. *Eur. J. Biochem.* **257**, 149–153.
62 Hajjar E (2007) *Molecular modeling of the substrate specificity and of the membrane anchoring of Proteinase 3*.
63 Introduction to Homology Modeling *Basic Concepts in Structural Bioinformatics*.
64 Consortium TU (2014) Activities at the Universal Protein Resource (UniProt). *Nucl. Acids Res.* **42**, D191–D198.
65 McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP & Lopez R (2013) Analysis Tool Web Services from the EMBL-EBI. *Nucl. Acids Res.* **41**, W597–W600.
66 Bai D & Wang AH (2014) Extracellular domains play different roles in gap junction formation and docking compatibility. *Biochem. J.* **458**, 1–10.
67 Hall T (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**, 95–98.
68 SAVES. The Structure Analysis and Verification Server version 4, Molecular Biology Institute at the University of California, Los Angeles
69 SAVES
70 Nielsen M, Lundegaard C, Lund O & Petersen TN (2010) CPHmodels-3.0—remote homology modeling using structure-guided sequence profiles. *Nucl. Acids Res.* **38**, W576–W581.
71 *The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.*
72 Bioinformatics Software
73 Guex N & Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723.
74 Humphrey W, Dalke A & Schulten K (1996) VMD: visual molecular dynamics. *J Mol Graph* **14**, 33–38, 27–28.
75 Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, Chipot C, Skeel RD, Kalé L & Schulten K (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802.

76 Mackerell AD (2004) Empirical force fields for biological macromolecules: overview and issues. *J Comput Chem* **25**, 1584–1604.
77 Adcock SA & McCammon JA (2006) Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem Rev* **106**, 1589–1615.
78 MacKerell AD, Brooks B, Brooks CL, Nilsson L, Roux B, Won Y & Karplus M (2002) CHARMM: The Energy Function and Its Parameterization. In *Encyclopedia of Computational Chemistry* John Wiley & Sons, Ltd.
79 CHARMM, http://www.charmm.org/
80 Carloni P Simulation of Biomolecules. *German Research School for Simulation Sciences*.
81 Steinbach PJ (2010) Introduction to Macromolecular Simulation. .
82 NAMD, http://www.ks.uiuc.edu/Research/namd/
83 Kozakov D, Beglov D, Bohnuud T, Mottarella SE, Xia B, Hall DR & Vajda S (2013) How Good is Automated Protein Docking? *Proteins* **81**, 2159–2166.
84 Kozakov D, Brenke R, Comeau SR & Vajda S (2006) PIPER: An FFT-based protein docking program with pairwise potentials. *Proteins* **65**, 392–406.
85 Comeau SR, Gatchell DW, Vajda S & Camacho CJ (2004) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. *Bioinformatics* **20**, 45–50.
86 Comeau SR, Gatchell DW, Vajda S & Camacho CJ (2004) ClusPro: a fully automated algorithm for protein–protein docking. *Nucleic Acids Res* **32**, W96–W99.
87 Sobolev V, Eyal E, Gerzon S, Potapov V, Babor M, Prilusky J & Edelman M (2005) SPACE: a suite of tools for protein structure prediction and analysis based on complementarity and environment. *Nucleic Acids Res* **33**, W39–W43.
88 Vriend G (1990) WHAT IF: a molecular modeling and drug design program. *J Mol Graph* **8**, 52–56, 29.
89 My PyMOL Script repository, http://pldserver1.biochem.queensu.ca/~rlc/work/pymol/
90 Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS & Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **30**, 2785–2791.
91 Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK & Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19**, 1639–1662.
92 Huey R, Morris GM, Olson AJ & Goodsell DS (2007) A semiempirical free energy force field with charge-based desolvation. *J. Comput. Chem.* **28**, 1145–1152.
93 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN & Bourne PE (2000) The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
94 Guex N & Peitsch MC (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**, 2714–2723.
95 Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L & Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucl. Acids Res.* **42**, W252–W258.
96 Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY & Prokopenko VV (2005) Virtual computational chemistry laboratory--design and description. *J. Comput. Aided Mol. Des.* **19**, 453–463.
97 VCCLAB, Virtual Computational Chemistry Laboratory, http://www.vcclab.org (2005) *VCCLAB, Virtual Computational Chemistry Laboratory*.
98 Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z & Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–672.
99 Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B & Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **36**, D901–906.
100 ArgusLab, Thompson, M. A., http://www.arguslab.com/ *http://www.arguslab.com/*.
101 Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS & Olson AJ (2009) AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **30**, 2785–2791.
102 *Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 4.5, San Diego: Dassault Systèmes, 2015*

103 Scopes RK (2001) Enzyme Activity and Assays. In *eLS* John Wiley & Sons, Ltd.
104 Biosciences A (2001) *Protein Purification: Handbook* Amersham Pharmacia Biotech.
105 AC03640923 A (2002) *Gel Filtration: Principles and Methods* Amersham Biosciences AB.
106 Strelow J, Dewe W, Iversen PW, Brooks HB, Radding JA, McGee J & Weidner J (2004) Mechanism of Action assays for Enzymes. In *Assay Guidance Manual* (Sittampalam GS, Gal-Edd N, Arkin M, Auld D, Austin C, Bejcek B, Glicksman M, Inglese J, Lemmon V, Li Z, McGee J, McManus O, Minor L, Napper A, Riss T, Trask OJ, & Weidner J, eds), Eli Lilly & Company and the National Center for Advancing Translational Sciences, Bethesda (MD).
107 Neubig RR, Spedding M, Kenakin T & Christopoulos A (2003) International Union of Pharmacology Committee on Receptor Nomenclature and Drug Classification. XXXVIII. Update on Terms and Symbols in Quantitative Pharmacology. *Pharmacol Rev* **55**, 597–606.
108 Shoolingin-Jordan PM & Leadbeater R (1997) Coupled assay for uroporphyrinogen III synthase. In (Enzymology B-M in, ed), pp. 327–336. Academic Press.
109 Fersht A (1999) *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* W. H. Freeman.
110 Veitinger S (2011) The Patch-Clamp Technique. *Leica Sciencelab*.
111 Nakagawa S, Gong X-Q, Maeda S, Dong Y, Misumi Y, Tsukihara T & Bai D (2011) Asparagine 175 of Connexin32 Is a Critical Residue for Docking and Forming Functional Heterotypic Gap Junction Channels with Connexin26. *J. Biol. Chem.* **286**, 19672–19681.
112 Barrio LC, Suchyna T, Bargiello T, Xu LX, Roginski RS, Bennett MV & Nicholson BJ (1991) Gap junctions formed by connexins 26 and 32 alone and in combination are differently affected by applied voltage. *Proc. Natl. Acad. Sci. U.S.A.* **88**, 8410–8414.
113 Jenkins R (2006) X-Ray Techniques: Overview. In *Encyclopedia of Analytical Chemistry* John Wiley & Sons, Ltd.
114 Putnam CD, Hammel M, Hura GL & Tainer JA (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics* **40**, 191–285.
115 Smyth MS & Martin JHJ (2000) x Ray crystallography. *Mol Pathol* **53**, 8–14.
116 Petoukhov MV, Franke D, Shkumatov AV, Tria G, Kikhney AG, Gajda M, Gorba C, Mertens HDT, Konarev PV & Svergun DI (2012) New developments in the *ATSAS* program package for small-angle scattering data analysis. *Journal of Applied Crystallography* **45**, 342–350.
117 Konarev PV, Volkov VV, Sokolova AV, Koch MHJ & Svergun DI (2003) *PRIMUS* : a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography* **36**, 1277–1282.
118 Svergun D, Barberato C & Koch MHJ (1995) *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography* **28**, 768–773.
119 Svergun DI (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886.
120 Kozin MB & Svergun DI (2001) Automated matching of high- and low-resolution structural models. *Journal of Applied Crystallography* **34**, 33–41.
121 Kabsch W (2010) XDS. *Acta Crystallogr D Biol Crystallogr* **66**, 125–132.
122 Vagin A & Teplyakov A (1997) *MOLREP* : an Automated Program for Molecular Replacement. *Journal of Applied Crystallography* **30**, 1022–1025.
123 Emsley P & Cowtan K (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132.
124 Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F & Vagin AA (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr D Biol Crystallogr* **67**, 355–367.
125 Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AGW, McCoy A, McNicholas SJ, Murshudov GN, Pannu NS, Potterton EA, Powell HR, Read RJ, Vagin A & Wilson KS (2011) Overview of the *CCP* 4 suite and current developments. *Acta Crystallographica Section D Biological Crystallography* **67**, 235–242.
126 Bruker Corp. Nuclear Magnetic Resonance (NMR) <web page>. .

127 Hoffman R (2013) What is NMR? *The NMR Lab, Institute of Chemistry, Hebrew University*.
128 Tadeo Campillo X (2010) *Structural basis for the aminoacid compositions of halophilic proteins and its relationship with enzymatic activity and stability*.
129 Kessler H, Gehrke M & Griesinger C (1988) Two-Dimensional NMR Spectroscopy: Background and Overview of the Experiments [New Analytical Methods (36)]. *Angew. Chem. Int. Ed. Engl.* **27**, 490–536.
130 Cavanagh J, Fairbrother WJ, Palmer III AG, Rance M & Skelton NJ (2007) Chapter 3 - Experimental aspects of NMR spectroscopy. In *Protein NMR Spectroscopy (Second Edition)* (Cavanagh J, Fairbrother WJ, Palmer AG, Rance M, & Skelton NJ, eds), pp. 114–270. Academic Press, Burlington.
131 Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J & Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293.
132 Zonta F, Polles G, Sanasi MF, Bortolozzi M & Mammano F (2013) The 3.5 ångström X-ray structure of the human connexin26 gap junction channel is unlikely that of a fully open channel. *Cell Commun. Signal* **11**, 15.
133 Gong X-Q, Nakagawa S, Tsukihara T & Bai D (2013) A mechanism of gap junction docking revealed by functional rescue of a human-disease-linked connexin mutant. *J Cell Sci* **126**, 3113–3120.
134 Nagy J i., Ionescu A-V, Lynn B d. & Rash J e. (2003) Coupling of astrocyte connexins Cx26, Cx30, Cx43 to oligodendrocyte Cx29, Cx32, Cx47: Implications from normal and connexin32 knockout mice. *Glia* **44**, 205–218.
135 Richard G, Brown N, Ishida-Yamamoto A & Krol A (2004) Expanding The Phenotypic Spectrum of Cx26 Disorders: Bart–Pumphrey Syndrome is Caused by a Novel Missense Mutation in GJB2. *J Investig Dermatol* **123**, 856–863.
136 Akiyama M, Sakai K, Arita K, Nomura Y, Ito K, Kodama K, McMillan JR, Kobayashi K, Sawamura D & Shimizu H (2007) A Novel GJB2 Mutation p.Asn54His in a Patient with Palmoplantar Keratoderma, Sensorineural Hearing Loss and Knuckle Pads. *J Invest Dermatol* **127**, 1540–1543.
137 Alexandrino F, de Oliveira CA, Magalhães RF, Florence MEB, de Souza EM & Sartorato EL (2009) Connexin mutations in Brazilian patients with skin disorders with or without hearing loss. *Am. J. Med. Genet. A* **149A**, 681–684.
138 Tsukada K, Nishio S, Usami S & Deafness Gene Study Consortium (2010) A large cohort study of GJB2 mutations in Japanese hearing loss patients. *Clin. Genet.* **78**, 464–470.
139 Feng Y, He C feng, Xiao J, Fang J, Zhao S, Tian X 'e, Mei L, Xia K & Tang X (2002) [An analysis of a large hereditary postlingually deaf families and detecting mutation of the deafness genes]. *Lin Chuang Er Bi Yan Hou Ke Za Zhi* **16**, 323–325.
140 Oshima A (2014) Structure and closure of connexin gap junction channels. *FEBS Letters* **588**, 1230–1237.
141 Kwon T, Roux B, Jo S, Klauda JB, Harris AL & Bargiello TA (2012) Molecular Dynamics Simulations of the Cx26 Hemichannel: Insights into Voltage-Dependent Loop-Gating. *Biophys J* **102**, 1341–1351.
142 Oh S, Rivkin S, Tang Q, Verselis VK & Bargiello TA (2004) Determinants of gating polarity of a connexin 32 hemichannel. *Biophys. J.* **87**, 912–928.
143 Bargiello TA, Tang Q, Oh S & Kwon T (2012) Voltage-dependent conformational changes in connexin channels. *Biochim Biophys Acta* **1818**, 1807–1822.
144 Delmar M, Coombs W, Sorgen P, Duffy HS & Taffet SM (2004) Structural bases for the chemical regulation of Connexin43 channels. *Cardiovascular Research* **62**, 268–275.
145 Locke D, Kieken F, Tao L, Sorgen PL & Harris AL (2011) Mechanism for modulation of gating of connexin26-containing channels by taurine. *J Gen Physiol* **138**, 321–339.
146 Orthmann-Murphy JL, Enriquez AD, Abrams CK & Scherer SS (2007) Loss-of-function connexin47 mutations cause Pelizaeus-Merzbacher-like disease. *Mol Cell Neurosci* **34**, 629–641.
147 Biancheri R, Rosano C, Denegri L, Lamantea E, Pinto F, Lanza F, Severino M & Filocamo M (2013) Expanded spectrum of Pelizaeus–Merzbacher-like disease: literature revision and description of a novel GJC2 mutation in an unusually severe form. *Eur J Hum Genet* **21**, 34–39.
148 Xin L & Bai D (2013) Functional roles of the amino terminal domain in determining biophysical properties of Cx50 gap junction channels. *Front Physiol* **4**.

149 Fortian A, González E, Castaño D, Falcon-Perez JM & Millet O (2011) Intracellular rescue of the uroporphyrinogen III synthase activity in enzymes carrying the hotspot mutation C73R. *J. Biol. Chem.* **286**, 13127–13133.
150 Blouin J-M, Duchartre Y, Costet P, Lalanne M, Ged C, Lain A, Millet O, de Verneuil H & Richard E (2013) Therapeutic potential of proteasome inhibitors in congenital erythropoietic porphyria. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18238–18243.
151 Bdira F ben, González E, Pluta P, Laín A, Sanz-Parra A, Falcon-Perez JM & Millet O (2014) Tuning intracellular homeostasis of human uroporphyrinogen III synthase by enzyme engineering at a single hotspot of congenital erythropoietic porphyria. *Hum. Mol. Genet.*, ddu298.
152 Jordan PM & Woodcock SC (1991) Mutagenesis of arginine residues in the catalytic cleft of Escherichia coli porphobilinogen deaminase that affects dipyrromethane cofactor assembly and tetrapyrrole chain initiation and elongation. *Biochem. J.* **280 (Pt 2)**, 445–449.
153 Warren MJ & Jordan PM (1988) Investigation into the nature of substrate binding to the dipyrromethane cofactor of Escherichia coli porphobilinogen deaminase. *Biochemistry* **27**, 9020–9030.
154 Leeper FJ & Rock M (1996) Interaction of analogues of porphobilinogen with porphobilinogen deaminase. *J. Chem. Soc., Perkin Trans. 1*, 2643–2649.
155 Araujo LS, Lombardo ME & Batlle AM (1994) Inhibition of porphobilinogenase by porphyrins in Saccharomyces cerevisiae. *Int. J. Biochem.* **26**, 1377–1381.
156 Jones RM & Jordan PM (1994) Purification and properties of porphobilinogen deaminase from Arabidopsis thaliana. *Biochem J* **299**, 895–902.
157 Grandchamp B, Phung N, Grelier M & Nordmann Y (1976) The spectrophotometric determination of uroporphyrinogen I synthetase activity. *Clin. Chim. Acta* **70**, 113–118.
158 Magnussen CR, Levine JB, Doherty JM, Cheesman JO & Tschudy DP (1974) A red cell enzyme method for the diagnosis of acute intermittent porphyria. *Blood* **44**, 857–868.
159 Lim CK, Li FM & Peters TJ (1988) High-performance liquid chromatography of porphyrins. *J. Chromatogr.* **429**, 123–153.
160 Anderson PM & Desnick RJ (1980) Purification and properties of uroporphyrinogen I synthase from human erythrocytes. Identification of stable enzyme-substrate intermediates. *J. Biol. Chem.* **255**, 1993–1999.
161 Jordan PM & Woodcock SC (1991) Mutagenesis of arginine residues in the catalytic cleft of Escherichia coli porphobilinogen deaminase that affects dipyrromethane cofactor assembly and tetrapyrrole chain initiation and elongation. *Biochem. J.* **280 (Pt 2)**, 445–449.
162 Jordan PM, Thomas SD & Warren MJ (1988) Purification, crystallization and properties of porphobilinogen deaminase from a recombinant strain of Escherichia coli K12. *Biochem. J.* **254**, 427–435.
163 Azim N, Deery E, Warren MJ, Wolfenden BAA, Erskine P, Cooper JB, Coker A, Wood SP & Akhtar M (2014) Structural evidence for the partially oxidized dipyrromethene and dipyrromethanone forms of the cofactor of porphobilinogen deaminase: structures of the Bacillus megaterium enzyme at near-atomic resolution. *Acta Crystallogr D Biol Crystallogr* **70**, 744–751.
164 Shoolingin-Jordan PM, Warren MJ & Awan SJ (1996) Discovery that the assembly of the dipyrromethane cofactor of porphobilinogen deaminase holoenzyme proceeds initially by the reaction of preuroporphyrinogen with the apoenzyme. *Biochem J* **316**, 373–376.
165 Roberts A, Gill R, Hussey RJ, Mikolajek H, Erskine PT, Cooper JB, Wood SP, Chrystal EJT & Shoolingin-Jordan PM (2013) Insights into the mechanism of pyrrole polymerization catalysed by porphobilinogen deaminase: high-resolution X-ray studies of the Arabidopsis thaliana enzyme. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 471–485.
166 Matthews BW (1968) Solvent content of protein crystals. *J. Mol. Biol.* **33**, 491–497.
167 Kantardjieff KA & Rupp B (2003) Matthews coefficient probabilities: Improved estimates for unit cell contents of proteins, DNA, and protein–nucleic acid complex crystals. *Protein Sci* **12**, 1865–1871.
168 Ellis RJ & Minton AP (2006) Protein aggregation in crowded environments. *Biol. Chem.* **387**, 485–497.
169 Shoolingin-Jordan PM & Leadbeater R (1997) Coupled assay for uroporphyrinogen III synthase. *Meth. Enzymol.* **281**, 327–336.

170 Schubert HL, Raux E, Matthews M a. A, Phillips JD, Wilson KS, Hill CP & Warren MJ (2002) Structural diversity in metal ion chelation and the structure of uroporphyrinogen III synthase. *Biochem. Soc. Trans.* **30**, 595–600.
171 Schubert HL, Phillips JD, Heroux A & Hill CP (2008) Structure and mechanistic implications of a uroporphyrinogen III synthase-product complex. *Biochemistry* **47**, 8648–8655.
172 Cunha L, Kuti M, Bishop DF, Mezei M, Zeng L, Zhou M-M & Desnick RJ (2008) Human uroporphyrinogen III synthase: NMR-based mapping of the active site. *Proteins* **71**, 855–873.
173 Ortega G, Castaño D, Diercks T & Millet O (2012) Carbohydrate Affinity for the Glucose–Galactose Binding Protein Is Regulated by Allosteric Domain Motions. *J. Am. Chem. Soc.* **134**, 19869–19876.

PUBLICATIONS

7

Paula Pluta, Oscar Millet, Luis Barrio, Structural analysis of the intercellular gap junctions produced by Cx26, Cx32 and Cx47, manuscript in preparation

Fredj ben Bdira, Esperanza González, **Paula Pluta**, Ana Laín, Arantza Sanz-Parra, Juan Manuel Falcón-Pérez, Oscar Millet, Human Molecular Genetics, Tuning intracellular homeostasis of human uroporphyrinogen III synthase by enzyme engineering at a single hotspot of congenital erythropoietic porphyria, 2014, doi: 10.1093/hmg/ddu298

Chahrazade Kantari, Arnaud Millet, Julie Gabillet, Eric Hajjar, Torben Broemstrup, **Paula Pluta**, Nathalie Reuter, Véronique Witko-Sarsat, Journal of Leukocyte Biology, Molecular analysis of the membrane insertion domain of proteinase 3, the Wegener's autoantigen, in RBL cells: implication for its pathogenic activity, 2011, 90(5):941-50, doi: 10.1189/jlb.1210695

ACKNOWLEDGMENTS

8

Funding & Grants

This thesis would not have been possible without the financial support of:

FUNDACIÓN RAMÓN ARECES

CIC bioGUNE
Biozientzietako Ikerkuntza Kooperatiboko Zentroa
Centro de Investigación Cooperativa en Biociencias

Research for this thesis was carried out at the Structural Biology Unit at the CIC bioGUNE institute.

EMBO Fellow

Short-Term Fellowship Ref: ASTF 46 - 2015

INDIEGOGO

And those who supported “Paula & the proteins” project:
Abraham Lopez Asamar, Alejandro Cantero Fuertes, Arvind Suresh, Carlos Besga Arcauz, Elizaveta Katorcha, Eneko Martin Ramila, Ewa Gubb, Ewa Kubala, Gabriel Ortega Quintanilla, Ganeko Bernardo Seisdedos, Gorka Lasso Cabrera, Javier Salcedo Conde, Jonatan and Begoña Ruiz Orbezua, José Ferreira, Leonor Guedes Silva, Maria Samalikova, Marie Miglianico, Panchami Prabhakaran, Sandipan Chowdhury, Sebastiaan van Liempd, Sivanandam Veeramuthu Natarajan, Wiesław Pluta, Wojciech Prusicki

Acknowledgements * Agradecimientos * Podziękowania

I would like to sincerely thank:

- Dr. Oscar Millet for giving me the chance to be a part of his group and supervising all my research.
- Dr Adriana Rojas and Dr. Luis Barrio who have contributed to my projects – protein-ligand and protein-protein parts, respectively – by training my crystallization skills and providing experimental data.
- All members of the CIC bioGUNE institute who shared with me space and knowledge. Especially the members of the group – current and former – who, more than lab-mates, were lab-friends to me.
- All members of the Queen Mary Biochemistry Laboratory for their hospitality and for making my short internship not only productive but also enjoyable. Above all Prof. Jon Cooper and Prof. Richard Pickersgill who made my visit possible.

Muchas gracias:

- A Bego & Jon por todos los tiempos verbales que hemos estudiado y todas las olas que hemos cogido. Por todos los amigos que me han presentado y todos los sitios donde me han traido.
- A Eneko por la mariposa
- A mi "Hermano Caramelo" por compartir conmigo el escritorio, miles de ideas y toda la comida india que me ha preparado.
- A Alberto, Eder, Javi & cuadrilla, Lander & cuadrilla por acceptar los giris entre sus amigos. Al final, los de Bilbao, nacemos donde queremos!
- A todos con quien estaba compartiendo la realidad bilbaina mágica; en el trabajo, en las casas de Ajuriaguerra y de Ercilla , en todos los centros deportivos, por las calles y por los bares.

Dziękuję bardzo:

- Całej mojej Rodzinie, a w szczególności Rodzicom – za to, że od 30 lat wpierają moje pomysły i plany, nawet jeśli części nie akceptują.
- Michałowi – za to, że jest.
- Dr Barbarze A Zarzyckiej – za to, że przetrwała ze mną nie tylko doktorat, ale nawet chemię fizyczną; za szczerość, inspirację i nieziemską wyrozumiałość.

- Adze, Emi i Ali – za to, że pomimo przemawiającego na moją niekorzyść faktu dzielącej nas odległości i wynikających z tego przypadków niedopełnienia lub zaniedbania obowiązków przyjaciółki podjęły decyzję o przedłużeniu mi karty członkowskiej w Klubie. Językiem nieprawniczej większości społeczeństwa – za przyjaźń, która prztrwała próbę czasu i kilometrów! Za to samo dziękuję Arturowi B., Piotrkowi B., Radkowi S. i Wojtkowi P.

My deepest gratitude
to everyone who finds my research interesting
and to everyone who ever has shown me love, support and encouragement.

… because those are the interactions that matter most.

CONTENTS

Universidad del PaisVasco / Euskal Herriko Universitatea