Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 3541 – 3546

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimization and Computing (ICMOC-2012)

A Study on Protein (*P-glycoprotein*) Homology Detection using Hidden Markov Model

Anurag Pal*, Debahuti Mishra, Shruti Mishra, Sandeep Kumar Satapathy and Kaberi Das

Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India

Abstract

P-glycoprotein (P-gp) is an ATP-dependent transport protein. It is selectively expressed at entry points of xenobiotics where, acting as an efflux pumps which prevents the entry into sensitive organs. This protein also plays a key role in the absorption and blood-brain barrier penetration of many drugs, while it's over expression in cancer cells has been linked to multidrug resistance (MDR) in tumors. In P-gp, the MDR is the principal mechanism by which many tumor cells develop resistance to chemotherapy drugs. The tumor begins to grow again; chemotherapy may fail because the remaining tumor cells are now resistant. Flexible receptor docking has been used to develop new prediction algorithm for P-gp binding specificity. Protein homology detection and sequence alignment are at the basis of protein structure prediction, function prediction and evolutionary analysis. Hidden Markov Models (HMMs) are usually used for statistical modeling, database searching, and multiple alignments of protein families and protein domains. In this paper, a detailed survey on P-gp (its function and structure), HMM (its function and structure with respect to P-gp) and MDR has been given.

Keywords: P-glycoprotein (P-gp); Multi Drug Resistance (MDR); Multiple Alignments; Protein Homology; Hidden Markov Model (HMMs)

1. Introduction

P-glycoprotein [1] (permeability glycoprotein, abbreviated as P-gp or Pgp) is a well-characterized ABC-transporter of the MDR/TAP subfamily. P-gp is also called ABCB1, ATP-binding cassette sub-family B member 1, MDR1, and PGY1. P-glycoprotein has also recently been designated CD243 (cluster of differentiation 243). In humans, P-glycoprotein is encoded by the ABCB1 gene. A research team at the Scripps Research Institute has obtained the first glimpse of a protein that keeps certain substances, including many drugs out of cells [8]. The protein, called P-glycoprotein, is one of the main reasons for which cancer cells are resistant to chemotherapy drugs. Understanding its structure may help scientists design more effective drugs. Resistance to therapy has been correlated to the presence of at least two molecular pumps in tumor-cell membranes that actively expel chemotherapy drugs from the interior. This allows tumor cells to avoid the toxic effects of the drug or molecular processes within the nucleus or the cytoplasm. The two pumps commonly found to confer chemo resistance in cancer are P-gp and the so-called multidrug resistance–associated protein (MRP). Because of their function and importance, they are

* Tel.: +91-9437668393; fax: 91-674-2351880.
E-mail address: anuragom2000@yahoo.com

doi:10.1016/j.proeng.2012.06.408

the targets of several anticancer efforts [9]. In this paper, a detailed survey report is provided on the P-gp proteins stating about its function and structures in correlation to cancer cells. Also, a discussion on HMM model regarding its function and application towards P-gp and its use computational biology is given. The layout of the paper is as follows: section 2 discusses the preliminary concepts of proteins and glycoprotein, MDR, functions of P-gp, structure of P-gp and mechanism of P-gp. Section 3 deals with general definition of the HMM model, its function and its role in computational biology and finally section 4 deals about the conclusion and its future directions.

2. P-Glycoprotein

2.1 Protein

Proteins are biochemical compounds consisting of one or more polypeptides typically folded into a globular or fibrous form, facilitating a biological functions [1-2]. A polypeptide is a single linear polymer chain of amino acids bonded together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. Proteins are an important class of biological macromolecules present in all organisms. All proteins are polymers of amino acids; classified by their physical size, proteins are nanoparticles (definition: 1–100 nm)[2][20]. Each protein polymer – also known as a polypeptide – consists of a sequence of 20 different L-α-amino acids, also referred to as residues. For chains under 40 residues the term peptide is frequently used instead of protein. To be able to perform their biological function, proteins fold into one or more specific spatial conformations, driven by a number of non-covalent interactions such as hydrogen bonding, ionic interactions, Van Der Waals forces, and hydrophobic packing. To understand the functions of proteins at a molecular level, it is often necessary to determine their three-dimensional structure [1-2].

Glycoproteins also form connective tissues such as collagen. They are usually found in gastrointestinal mucus secretions and are used as protective agents and lubricants [1-2][19-20]. They are also found abundantly in the blood plasma where they serve many functions. The diverse function of glycoprotein is a direct result of their structure. These macromolecules are composed of a peptide chain with one or more carbohydrate moieties. There are two broad categories of glycoprotein structure. The carbohydrates are either linked N-glycosidically or O-glycosidically to their constituent protein. Within these broader categories, there can be fine structural differences which account for the large diversity of functions among glycoprotein. Controlling of glycoproteins is achieved through synthesis and degradation. Those processes are controlled by very specific enzymes.

2.2 Multi Drug Resistance

MDR is a condition enabling a disease-causing organism to resist distinct drugs or chemicals of a wide variety of structure and function targeted at eradicating the organism [8]. Organisms that display multidrug resistance can be pathologic cells, including bacterial and neoplastic (tumor) cells. Many different bacteria now exhibit MDR, including staphylococci, enterococci, gonococci, streptococci, salmonella, Mycobacterium tuberculosis, and others [8-9]. In addition, some resistant bacteria are able to transfer copies of DNA that codes for a mechanism of resistance to other bacteria, thereby conferring resistance to their neighbors, which then are also able to pass on the resistant gene. This process is called horizontal gene transfer [1].

2.3 Functions of P-gp

P-gp, the most extensively studied ATP-binding cassette (ABC) transporter, functions as a biological barrier by extruding toxins and xenobiotics out of cells. In vitro and in vivo studies have demonstrated that P-gp plays a significant role in drug absorption and disposition [10]. Because of its localization, P-gp appears to have a greater impact on limiting cellular uptake of drugs from blood circulation into brain and from intestinal lumen into epithelial cells than on enhancing the excretion of drugs out of hepatocytes and renal tubules into the adjacent luminal space. However, the relative contribution of intestinal P-gp to overall drug absorption is unlikely to be quantitatively important unless a very small oral dose is given, or the dissolution and diffusion rates of the drug are very slow [11]. This is because P-gp transport activity becomes saturated by high concentrations of drug in the intestinal lumen. Because of its importance in pharmacokinetics, P-gp transport screening has been incorporated into the drug discovery process, aided by the availability of transgenic mdr knockout mice and in vitro cell systems [12].

While the role of P-gp to keep potentially harmful compounds out of the brain may seem advantageous, it also poses a problem in the pharmacologic treatment of diseases such as brain tumors and brain infections in AIDS patients .The drugs intended to treat diseases affecting the brain are often P-gp substrates; this results in their expulsion even before they have a chance to treat the underlying pathology. Increasing the concentration of the drug to achieve entry into the brain to circumvent P-gp's action might present problems such as systemic toxicity. The blood-brain barrier is composed of specialized endothelial cells that prevent various substances from entering the brain. P-gp is an important component of this barrier and is present in high concentration on the apical surface of these endothelial cells [13]. It is an ATP-dependent transport protein thought to be involved in extruding a variety of structurally unrelated compounds and preventing their accumulation within the brain. It is known that P-gp impedes the entry of various drugs that are used in the treatment of central nervous system diseases. Understanding the structure and the function of P-gp will lead to the development of specific and selective P-gp inhibitors. Combined use of these inhibitors along with therapeutic agents could treat central nervous system diseases and result in improved clinical efficacy [14].

2.4 Structure of P-gp

A major factor contributing to drug resistance in cancer is the over-expression of P-gp, a plasma membrane ATP-binding cassette (ABC) drug efflux pump. Three-dimensional structural data with a resolution limit of ~8 Å have been obtained from two-dimensional crystals of P-glycoprotein trapped in the nucleotide-bound state. Each of the two transmembrane domains of P-glycoprotein consists of six long α-helical segments. Five of the α-helices from each transmembrane domain are related by a pseudo-2-fold symmetry, whereas the sixth breaks the symmetry [15]. The two α-helices positioned closest to the (pseudo-) symmetry axis at the center of the molecule appear to be kinked. A large loop of density at the extracellular surface of the transporter is likely to correspond to the glycosylated first extracellular loop, whereas two globular densities at the cytoplasmic side correspond to the hydrophilic, nucleotide-binding domains. This is the first three-dimensional structure for an intact eukaryotic ABC transporter. Comparison with the structures of two prokaryotic ABC transporters suggests significant differences in the packing of the transmembrane α-helices within this protein family. P-gp detoxifies cells by exporting hundreds of chemically unrelated toxins but has been implicated in MDR in the treatment of cancers. Substrate promiscuity is a hallmark of P-gp activity, thus a structural description of poly-specific drug-binding is important for the rational design of anticancer drugs and MDR inhibitors [2].

2.5 Mechanism of P-gp

Mechanism of P-glycoprotein (Pgp) using single-molecule fluorescence resonance energy transfer (FRET). Pgp, a member of the ATP binding cassette family of transport proteins, is found in the plasma membrane of animal cells where it is involved in the ATP hydrolysis driven export of hydrophobic molecules. When expressed in the plasma membrane of cancer cells, the transport activity of Pgp can lead to the failure of chemotherapy by excluding the mostly hydrophobic drugs from the interior of the cell [16].

Human P-gp (ABCB1) is a primary multidrug transporter located in plasma membranes that, utilizes the energy of ATP hydrolysis to pump toxic xenobiotics out of cells. P-gp employs a most unusual molecular mechanism to perform this drug transport function. Multidrug transporters are involved in mediating the failure of chemotherapy in treating several serious diseases. The archetypal multidrug transporter P-gp confers resistance to a large number of chemically and functionally unrelated anti-cancer drugs by mediating efflux from cancer cells. The ability to efflux such a large number of drugs remains a biological enigma and the lack of mechanistic understanding of the translocation pathway used by P-gp prevents rational design of compounds to inhibit its function.

3. Hidden Markov Models

3.1 General Definition

HMM [3] [5] is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (*hidden*) states. An HMM [6-7] can be considered as the simplest dynamic bayesian network. In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a *hidden* Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics[17].

3.2 HMMs in Computational Biology

Computational biology is motivated by newly available and abundant raw molecular datasets gathered from a variety of organisms. Though the availability of this data marks a new era in biological research, it alone does not provide any biologically significant knowledge. The goal of computational biology is then to elucidate additional information regarding protein coding, protein function and many other cellular mechanisms from the raw datasets. This new information is required for drug design, medical diagnosis, medical treatment and countless fields of research [4]. The majority of raw molecular data used in computational biology corresponds to sequences of nucleotides corresponding to the primary structure of DNA and RNA or sequences of amino acids corresponding to the primary structure of proteins. Therefore the problem of inferring knowledge from this data belongs to the broader class of sequence analysis problems. Two of the most studied sequence analysis problems are speech recognition and language processing [18].

Biological sequences have the same left-to-right linear aspect as sequences of sounds corresponding to speech and sequences of words representing language. Consequently, the major computational biology sequence analysis problems can be mapped to linguistic problems. A common linguistic metaphor in computational biology is that of protein family classification as speech recognition. The metaphor suggests interpreting different proteins belonging to the same family as different vocalizations of the same word. Another metaphor is gene finding in DNA sequences as the parsing of language into words and semantically meaningful sentences. It follows that biological sequences can be treated linguistically with the same techniques used for speech recognition and language processing [19].

3.3 HMM in P-glycoprotein

Understanding the structure of P-gp and its mechanism is crucial to the development of P-gp inhibitors and systemic drugs as well. While the role of P-gp to keep potentially harmful compounds out of the brain may seem advantageous, it also poses a problem in the pharmacologic treatment of diseases such as brain tumors and brain infections in AIDS patients. P-gp can extrude a variety of structurally diverse, toxic xenobiotic compounds from cells. HMM also used for detecting remote protein homologies. HMM which is a statistical model also used to analyze, to predict proper protein sequence and that can help in many ways to medical science and in the field of bio-computing. Many drugs used in clinical therapy are P-gp substrates, and the transporter is now increasingly recognized to play a central role in the absorption and disposition of many drugs, including chemotherapeutic agents. P-gp is a drug transporter [20] also helpful in kidney and liver problems.

4. Conclusion

HMM is a very efficient, effective and accurate model. This is also rich in mathematical approach. We can apply it in different sectors, mostly in bioinformatics for finding gene sequence, protein sequence analysis and many more to get an appropriate result. Mainly in case of P-gp, we can apply HMM to study or predict the proper structure of protein and we can get a very strong and effective result and a fruitful solution in case of tumor cells.

References

[1] Agarwal P and States DJ. Comparative accuracy of methods for protein-sequence similarity search. *Journal of Bioinformatics*; 2010; 14: 1: 40-47.

[2] Levchenko A, Mehta BM, Niu X, Kang G, Villafania L, Way D, Polycarpe D, Sadelain M, and Larson S.M. Intercellular transfer of P-glycoprotein mediates acquired multidrug resistance in tumor cells. *Proc. National Academy of Science*; 2009;102: 1933-1938

[3] Henderson J, Salzberg S and Fasman K. Finding Genes in DNA with a Hidden Markov Model. *Journal of Computational Biology*; 2010; 4: 2: 127-141.

[4] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. of the IEEE Transactions*; 2009; 7: 1: 257-284.

[5] Künsch HR, Barndorff-Nielsen OE, Cox DR and Klüppelberg C. State Space and Hidden Markov Models in Complex Stochastic Systems. *London, Chapman and Hall/CRC;* 2010: 109–173.

[6] Ding J, Shah SP. Robust hidden semi-markov modeling of arra CGH data. *Proc. of IEEE Int. Conf. on Bioinformatics and Biomedicine*; 201:1**1: **17**:** 603-608

[7] Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A. Statistical issues in the analysis of the array CGH data. *IEEE Bioinformatica Conferences*; 2008; 1: 1: 407-408

[8] *http://en.wikipedia.org/wiki/P-glycoprotein*

[9] *http://www.fluorosome.com/pdf/p-glycoprotein-multidrug-resistance-protein.pdf*

[10]Ramakrishnan P. The Role of P-glycoprotein in thr Blood-Brain Barrier. *Einstein Journal of Biology and Medicine*; 2003; 19: 1:160-165.

[11]*http://www.ncbi.nlm.nih.gov/pubmed/8100632*

[12]Giacomini KM, Huang SM, Tweedie DJ, Benet LZ, Brouwer KLR, Xiaoyan C. Membrane transporters in drug development. *Nature Reviews Drug Discovery*; 2010; 9: 1: 215-236.

[13]Langford D, Grigorian A, Hurford R, Adame A, Ellis RJ, Hansen L, Masliah E. Altered P-glycoprotein expression in AIDS patients with HIV encephalitis. *Journal of Neuropathology & Experimental Neurology*; 2004; 63: 10:1038-47.

[14]Kazantsev AG, Thompson LM. Therapeutic application of histone deacetylase inhibitors for central nervous system disorders. *Nature Reviews Drug Discovery*; 2008; 7: 1: 854-868.

[15]Aller SG, Yu J, Ward A, Weng Y, Chittaboina S, Zhuo R, Harrell PM, Trinh YT, Zhang Q, Urbatsch IL, Chang G. Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. *Science*; 2009; 323: 5922: 1718-22.

[16]Ernst S, Verhalen B, Zarrabi N, Wilkens S, Boersch M. Drug transport mechanism of P-glycoprotein monitored by single molecule fluorescence resonance energy transfer. *Quantitative Biology*; 2011.

[17]Fonzo VD, Pentini FA,Parisi V. Hidden Markov Model in Bioinformatics. *Current Bioinformatics*; 2007; 2: 49-61.

[18]*http://www.ncbi.nlm.nih.gov/pubmed/8107089*

[19]*http://people.binf.ku.dk/krogh/publications/pdf/KroghEtal94a.pdf*

[20]*http://www.cs.toronto.edu/~radford/ftp/sun-thesis.pdf*