CHAPTER 8

SELECTION USING SRLCS MODEL FOR PROTEIN COMPARATIVE MODELING

8.1. Overview of Enzymes

A classic approach in Biology, both organismal and cellular, is to compare morphologies in order to glean structural and functional commonalities. The comparative approach has also proved valuable on a molecular level. For example, phylogenetic comparisons of RNA sequences have led to the determination of conserved secondary and even tertiary structures, and comparisons of protein structures have led to classifications of families of protein folds [22]. The aim of comparing RNA and protein enzymes is to learn about fundamental physical and chemical principles of biological catalysis. A particular protein enzyme present in human may also be present in other organisms but not in identical form. It is of interest in biomedical field to know the presence of an enzyme in various organisms. Figure.8.1. shows the structures of the four different types of lysozyme from different organisms were aligned using VAST\textsuperscript{11} and displayed using Cn3D\textsuperscript{12}. [44, Pg 15 ch 2].

Enzymes are biological catalysts, mainly proteins, generated by an organism to speed up chemical reactions. Most bodily functions rely on enzymes; without them, normal healthy life is impossible. Enzymes can either launch a reaction or speed it up.


\textsuperscript{12} http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
The chemicals that are transformed with the help of enzymes are called substrate. In the absence of enzymes, these chemicals are called reactant. Efficiency of enzymes lies in transforming substrates into usable products at the rate of ten times per second. There are an estimated 75,000 different enzymes in the human body; these chemical reactions are performed at an amazing rate. On the other hand, in the absence of enzymes, reactants may take hundreds of years to convert into a usable product, if they are able to do so at all.

Figure 8.1. Structural alignment of goose lysozyme (PDB code 153L), chicken egg white lysozyme (3LZT), and lysozymes from E. coli bacteriophages λ (1AM7) and T4 (1L92)

This is why enzymes are crucial in the sustenance of life on earth. The absence of enzymes is responsible for many diseases. In humans, a tragic disease called phenylketonuria (PKU), which causes severe mental retardation and even death in
infants, is the result of the absence of one type of enzyme. Tay-Sachs disease is a similarly tragic result of an enzyme deficiency. It causes retardation, paralysis, and often death in early childhood when left untreated.

Our ability to alter enzymes by inhibiting their functioning abilities has resulted in hundreds of life saving drugs. One example is Penicillin, a well-known antibiotic that can cure syphilis, pneumonia, and other illnesses. Penicillin works by bonding to the active sites of the disease-causing bacteria’s enzymes, ultimately destroying the bacteria’s ability to survive and reproduce.

8.2. Enzyme Identification and Homology Modeling

The enzyme identification falls into the discipline of homology modeling. Generally, Homology modeling also known as “Comparative Modeling” can provide the Molecular Biologists and Biochemists with sufficient information about the spatial arrangement of important residues in the protein and that may guide the design of new experiments. Homology modeling refers to modeling a protein 3D structure using a known experimentally determined structure of a homologous protein as a template. A protein structure is always of great assistance in the study of protein function, dynamics, interactions with ligands and other proteins, and even within pharmaceutical industry in drug discovery and drug design. Comparative modeling is not a straight forward approach. It involves many steps like template identification, amino acid sequence alignment, alignment correction, backbone generation, generation of loops, side chain generation and optimization, ab-initio loop building, overall model optimization and model verification including quality.

Template identification is identifying a homologous template protein for a target protein i.e the specimen protein or enzyme in search. Comparative modeling uses the protein sequence of specimen and template proteins. The sequence alignment
of the specimen sequence with template sequences in the database will reveal the
degree of similarity amongst them. Normally, the database used for this purpose is
Homologues database. This sequence alignment gives an idea on the general features
of the protein family, degree of conservation, the consensus sequence, etc. Features
like insertions and deletions, the location of the active site residues help in alignment
correction of the modeling process.

8.2.1. Homology Modeling Categorization

The scholar categorizes homology modeling as two distinct categories.
1. Unknown Specimen - Uncategorized. Nothing is known about the specimen
   protein or enzyme.
2. Known specimen - The specimen is in known category, but looking for some
   more information.

8.2.1.1. Issues in Homology Modeling with Unknown Specimen

In this, the specimen is possibly obtained from some organism and the
ancestral origin of the organism may be unknown. Even if its ancestral relationships
are known, may be some traits are unexplainable. Hence proteins causing such traits
are to be related. In all there is so much curiosity to label the protein or protein
enzyme by identifying its homologues.

In this process of homology modeling, the researcher would be bothered about
the answers and relevance to the following questions:

- What can be modeled?
- How to find a template for homology modeling?
- What are the error sources in modeling?
- How to assess the quality of a homology model?
Overall, the end result can be achieved only after ascertaining answers to all the above questions and it is a lengthy process. Figure.8.2. shows the activity at the highest abstraction level.

![Diagram of protein family identification](image)

**Figure 8. 2. Identification of protein family of an unknown protein sequence**

### 8.2.1.2. Issues in Homology Modeling with Known Specimen

This is the scenario of a clinical laboratory, where the samples, origins and details are known. i.e., the sample is taken for a specific purpose, may be from a deceased patient. The treatment could be based on missing enzymes, shortfall of enzymes or malfunction etc. Therefore, the effort will be to know “how different is this specimen from the others belonging to the same category?” The closest or the farthest member from the family can be selected. Figure.8.3. shows the purpose of protein sequence comparison in known specimen case. This is like: “I know I have an apple but whether it is Kashmir apple, Washington apple or something else?”

133
Such a difference is represented with homology extent with its family members. For the reasoning process to start, it is enough to know the difference between the specimen and template. In many cases this is enough to direct the efforts in the right path.

Figure 8.3. Protein sequences comparison with a known specimen

### 8.3. Identification of Homology Extent

Comparative Modeling is one of the key methods to Protein Structure Prediction. Of the 8 steps to comparative modeling, alignment accuracy and template structure are two important factors that contribute to the goodness of the experiments to be carried out by Biologists. Template Identification is identifying a homologous template protein for a target protein (Specimen). This is done from a homologue database which consists of homologous sequences. The MSA of the Specimen sequence with template sequences in the database will reveal the degree of similarity amongst them. The scholar has already derived resource efficient SRLCS model for basic alignment. As an application of SRLCS Model, the scholar has proposed a method, for the selection of the set of template proteins from homologue database by calculating the “Homology extent” measure of the specimen.
8.4. SRLCS model combined with fuzzy membership to obtain the Homology Extent

8.4.1. Membership Calculator

SRLCS model can bring out the |LCS| between the specimen and template proteins taking into account the similarity, identity and length of the sequences. The block diagram for calculating the membership value is shown in Figure 8.4.

\[ p = \frac{|LCS|}{|Query|} \]  
\[ \text{(Eq.8.1)} \]

The ratio of the obtained Length of LCS and the length of target sequence is described as agreement factor \( p \) and is calculated as

The Fuzzy membership function \( M \) derives the degree of agreement between the target protein and the template proteins and is defined by the following formula:

\[
\text{Membership Function } M = \begin{cases} 
1 & p > 1.0 \\
0 & p < 0 \\
\frac{p}{1} & 0 < p < 1 
\end{cases} \]  
\[ \text{(Eq.8.2)} \]

The Membership Function \( M \) can take a value between 0 and 1. A value closer to 1 is expected to be a close homology and value closer to 0 is expected to be distant
or remote homology. This M value can provide a good basis to know the useful enzyme present in other animals.

8.4.2. Homology Extent Calculation

The template proteins are from homolog database. The specimen could be the one chosen by the investigator. The protein enzymes Myoglobin, Lysosomes and QNR interested the researcher (for the reasons explained in the following paragraphs) to bring out the usefulness of the SRLCS model. The SRLCS model combined with membership calculator yields the homology extent on a pairwise basis between the specimen and database sequences. Figure 8.5. represents the functional diagram for this activity.

![Diagram](image)

**Figure 8. 5. Homology extent calculation of specimen sequence with reference to template sequences**

8.5. Enzymes of Interest

Myoglobin is an iron containing pigment present in muscle. It combines with oxygen to form oxymyoglobin. Oxymyoglobin acts as a store of oxygen that can be
used during strenuous exercise. Each myoglobin molecule consists of a single polypeptide chain with a haem group which has an affinity for oxygen stronger than that of hemoglobin. It is this Myoglobin that gives the red color to the exposed muscles. This myoglobin is found in vertebrates and invertebrates.

Lysosomes are membrane-enclosed organelles that contain an array of enzymes capable of breaking down all types of biological polymers - proteins, nucleic acids, carbohydrates, and lipids. In their simplest form, lysosomes are visualized as dense spherical vacuoles, but they can display considerable variation in size and shape as a result of differences in the materials that have been taken up for digestion. Lysosomes thus represent morphologically diverse organelles defined by the common function of degrading intracellular material. They function as the digestive system of the cell, serving both to degrade material taken up from outside the cell and to digest obsolete components of the cell itself. It contains about 50 different degradative enzymes that can hydrolyze proteins, DNA, RNA, polysaccharides and lipids. Mutations in the genes that encode these enzymes are responsible for more than 30 different human genetic diseases which are called lysosomal storage diseases because undegraded material accumulates within the lysosomes of affected individuals. Most of these diseases result from deficiencies in single lysosomal enzymes.

Quinolones are broad-spectrum antibacterial agents, commonly used both in human and veterinary medicine. Their extensive use has been associated with raising level of quinolone resistance. Quinolone resistances (QNR) are chromosomally encoded. QNR protein is a 218-amino acid protein belonging to the pentapeptide-repeat family of proteins that protects DNA from quinolone binding to topoisomerases. A search for Putative QNR can provide the QNR reservoir in other organisms [49].
8.6. HOGENOM Data Base

HOGENOM\textsuperscript{13} [53] is a database of homologous genes from fully sequenced organisms (bacteria, archaean and eukarya), structured under ACNUC sequence database management system. It allows selecting sets of homologous genes among species and to visualize multiple alignments and phylogenetic trees. HOGENOM is particularly useful for comparative sequence analysis, phylogeny and molecular evolution studies. ACNUC is a retrieval system for the nucleotide and protein sequence databases GenBank, EMBL, UniProt/SWISS-PROT or NBRF-PIR and for many other databases following the same formats.

The PBIL (Pôle Bio-Informatique Lyonnais) World Wide Web server developed at the Laboratory of Biometry and Evolutionary Biology and the Institute of Biology and Chemistry of Proteins - is dedicated to Molecular Biology and Ecology. This server allows to browse through a number of general or specialized sequence databases such as GenBank, NBRF, EMBL, SWISSPROT /TrEMBL, Hovergen, NRSub and EMGLib. It also allows access for many analytical tools for nucleotide or protein sequence analysis. The PBIL server also offers many documents relevant to these fields of research as well as links with other information sources. This server is also devoted to the use of multivariate statistics in ecology.

8.7. Method and Results

Three proteins Myoglobin, Lysozyme and Quinolone resistance (QNR) are used for experiment. The sequences were taken from HOGENOM data base. These protein enzymes are found in most of the animals. The experiment was done individually on each of the three enzymes. In each case a Homosapien sequence is

\textsuperscript{13} This database is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version. A copy of the GNU General Public License is available at ftp://pbil.univ-lyon1.fr/pub/hogenom and http://www.gnu.org/licenses/.
used as template sample sequence and the rest used as homolog. SRLCS Model combined with fuzzy Membership calculator provided the homology extent results of the Homosapien enzyme to that of the homolog enzyme of other animals. This is displayed as a Graphical output along with membership value. This membership value is synonym to Homology extent.

8.7.1. Myoglobin Identification

Human myoglobin (HS22-37-PE42) is taken as reference and hence its membership value is referenced as 1.0. The closeness of the myoglobin of 14 other organisms with reference to the said human myoglobin is calculated in terms of the membership value and is shown in figure.8.6. It is observed that Horse, Mouse, Sperm Whale, Whale and Chicken which are all vertebrates have membership value > 0.77 implying the presence of close homologous Myoglobin protein. Other invertebrate organisms like Bacteria and single cell archa have poor homology or distant homology with reference to this human myoglobin protein.

Figure 8. 6. Membership information of Myoglobin HS22-37-PE42 Vs. other organisms
8.7.2. Lysosomes Identification

There were about 712 Lysosomes listed as homolog by HOGENOM protein database. HS10_33_PE14 is a Homosapien Lysosyme of length 194. When applied to SRLCS and then their membership with HS10-33PE is found to be very poor. This is because Lysosomes themselves are more than 50 varieties in an organism. And hence Inter organism comparison does not yield any closeness to the subject Lysosome found in human. Except for HS10-32, all others have membership value < 0.54. Figure.8.7. explains this result. This experiment was conducted with 3 different specimens. But nothing yielded any closeness to each other.

Similarly the membership of E-coli Lysosyme is found to be 1.0 only amongst its families. With other organisms the membership value is less than 0.56 which is a fact that lysosymes are widely varying. Figure.8.8. depicts the Ecoli result.
Figure 8.7 Membership information of Lysozyme HS 10-33PE14 Vs. other organisms
Figure 8.8. Membership information of Lysosyme ECO81_1_PE245 vs. other organism
8.7.3. Quinolone resistance (QNR)

Quinolones are anti bacterial agents useful for medicine. Search is on for this enzyme. QNR sequences available from HOGENOM are just 9. Once again Homosapien sequence **HS6_36_PE9** is taken as specimen query sequence. This is of length 276. The membership of QNR from other organisms is as in Figure 8.9. The membership is less than 0.5 with reference to all other organisms implying the scarcity of this enzyme in the animal world.

![Figure 8.9. Membership information of QNR HS 6-36 Vs. other organisms](image)

8.8. CONCLUSION

The SRLCS model combined with Fuzzy membership function is rightly identifying the similarity of specimen protein with template set of proteins from other organisms. This membership relationship so identified can rightly guide the Biologists in their investigation. Only a sample investigation is done with myoglobin, lysozyme and QNR proteins. However, such a membership finding can be done with any set of biosequences like DNA, Gene, RNA protein etc.