Sequence and Structural Analogies
between Glyceraldehyde-3-
Phosphate Dehydrogenase of Homo sapiens and the CysP Periplasmic Binding Protein from Escherichia coli

Garreesa S. Henry
Livingstone College

Faculty Mentor: Rufus M. Williamson
Livingstone College

ABSTRACT
Substrate-binding proteins are components of ATP-binding cassette transporters which capture their substrates in the periplasm of bacteria. These proteins subsequently deliver their bound ligands to membrane components of the transporters. Bacterial periplasmic substrate-binding proteins are characterized by a Venus flytrap motif, which is involved in the ability of the protein to capture their cognate substrates. Proteins with similarities to bacterial periplasmic binding proteins and having domains that resemble a Venus Flytrap have been also identified in eukaryotic cells. To further characterize eukaryotic proteins which resemble bacterial periplasmic binding proteins, bioinformatics techniques and procedures were used to survey the human genome. Results of amino acid sequence analyses reveal that glyceraldehyde-3-phosphate dehydrogenase displays a 49.6% amino acid sequence similarity to the CysP periplasmic binding protein, for sulfate, in Escherichia coli. Reliability estimates indicate that the likelihood that the two proteins belong to the same family of the proteins was approximately 99.6%. Secondary structural comparisons and homology modeling studies suggest that glyceraldehyde-3-phosphate dehydrogenase may contain a domain which resembles the Venus flytrap motif, that is common to periplasmic binding proteins. The results are interpreted to suggest that the two proteins are analogous.

ATP-binding cassette transporters are members of a protein superfamily that is one of the largest and most ancient families with representatives in all extant phyla from prokaryotes to humans. These transporters are composed of complexes that consist of two alpha-helical transmembrane domains, which form a translocation pathway, and two cytoplasmic domains, which power the transport reaction through binding and hydrolysis of adenosine triphosphate (ATP). The energy derived from hydrolysis is associated with the translocation of various substrates across membranes(1).

Periplasmic substrate-binding proteins are components of ATP-binding cassette transporters. These proteins bind their substrates, in the periplasm, selectively and with high affinity, and deliver bound ligands to the extracellular gate of the transmembrane domains within the transport complex. In bacteria, periplasmic binding proteins also play roles in chemotaxis(2), and intercellular communication(3) processes.
X-ray crystallographic studies have shown that periplasmic binding proteins consist of two large lobes that close around the bound ligand, resembling a Venus Flytrap\(^4\). The Venus Flytrap domain has been proposed to be encoded in an 8-residue motif\(^5\). Intriguingly, proteins with similarities to periplasmic binding proteins and having domains that resemble a Venus Flytrap have been identified in eukaryotic cells\(^1\).

The metabotropic receptor extracellular domain is a member of a family of structural domains linked to a variety of receptor types, including ionotropic glutamate receptors\(^6\). Both amino acid sequence and structural modeling studies have revealed that the metabotropic receptor extracellular domain is similar to bacterial periplasmic amino acid binding proteins\(^7\).

DING proteins, named for the presence of the amino acids, in their single letter abbreviation, DINGGGN, and located in the carboxyl termini of these proteins resemble periplasmic binding proteins. DING proteins are ubiquitous in living organisms\(^8\). Many DING proteins have been isolated in eukaryotes. They have been associated with very diverse biological activities, often in the context of possible signaling roles and are associated with both normal and pathological functions in mammals\(^8, 9\). Their phosphate-binding function suggests a role in biomineralization, and the ability to bind other ligands may be related to signal transduction in eukaryotes\(^9\). Additionally, a new family of receptor tyrosine kinases with a Venus Flytrap binding domain in insects and other invertebrates has been identified\(^10\). The Venus Flytrap domain is an ancient protein module present in multiple proteins and may represent a promising area for drug discovery research owing to the ability of proteins with the domain to capture molecules. It has been postulated that, in the process of evolution, genes for proteins containing the Venus Flytrap motif may have fused with genes with other proteins\(^8\).

To further characterize proteins which resemble bacterial periplasmic binding proteins in eukaryotes, a survey of proteins in the human genome was conducted. In this study, bioinformatics tools and procedures were used to identify glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) as a protein which resembles a periplasmic sulfate-binding protein in bacteria. Identification of additional proteins in eukaryotes with the Venus Flytrap domain could facilitate research in drug discovery, as well as, understanding of the molecular evolution of proteins.

**METHODS**

**Search Strategy.** A set of 70 amino acid sequences for bacterial periplasmic binding proteins was extracted from the Kyoto Genes and Genome database (KEGG)\(^11\). The KEGG database is a repository of information on the current knowledge of the genetics, biochemistry, and molecular biology of individual genes\(^11\). The set of amino acid sequences were compared and grouped by similarity into 7 distinct clusters. Members of a cluster frequently have shared substrate specificity (i.e., carbohydrates, amino acids, and metals). Consensus sequences were determined for each group by the methods of Brown, and Lai\(^12\) (data not shown). A consensus sequence is the representation of a sequence alignment in terms of the most frequently occurring amino acid residues found at each position. Conserved amino acid substitutions which are defined as replacement of an amino acid residue with another one with similar properties were considered. Also, semi-conserved substitution amino acid, defined as replacement of an amino acid residue with another that has similar steric
conformation, but does not share chemical properties were also considered. Amino acid replacement groups for conservative replacements, in one-letter code are STA, NEQK, NHQK, NDEQ, QHRK, MILV, MILF, HY, and FYW. Semi-conservative amino acid replacements are CSA, ATV, SAG, STNK, STPA, SGND, SNDEQK, NDEQHK, NEQHRK, FVLIM, and HFY. The one-letter code for amino acids is a useful way to display to represent amino acids.

Consensus sequences, along with, representative amino acid sequences for each group, were used as query sequences to search against *Homo sapiens* proteins data located in GenBank\(^{(13)}\) and SwissProt databases\(^{(14)}\). GenBank and SwissProt are comprehensive databases that contain publicly available nucleotide and/or protein sequences, as well as, functional information. We reasoned that it was unlikely to identify eukaryotic proteins that were homologous (i.e. proteins which share similar functions and common evolutionary ancestry) to bacterial periplasmic binding proteins. This is in part due to the evolutionary distance between bacteria and *Homo sapiens*, and the lack of the equivalent of a periplasmic space in eukaryotes. Thus, a search for analogous proteins was conducted. Analogous proteins share similar functions, but lack a common ancestral origin. Two approaches were used to identify putative similarities in the genome of *Homo sapiens*. Position-Specific Iterative Basic Local Alignment Searches (PSI-BLAST) were conducted to detect distant relationships between proteins\(^{(15)}\). In addition, to improve our ability to detect distantly related proteins that may not be identified by the PSI-BLAST algorithm, a second strategy was employed. The approach neglects the order of amino acid residues in a sequence, and uses properties of constituent amino acid instead to query the SwissProt database\(^{(16)}\). In this approach, a total of 144 individual characteristics for a given amino acid were used which included properties such as molecular weight, content of bulky residues, content of small residues, average hydrophobicity, and average charge.

**Protein Structural Studies.** Consensus secondary structure predictions, derived from several prediction algorithms, were performed using the Network Protein Sequence Analysis Tool\(^{(17)}\). Secondary structural elements are indicated by solid color coils (α-helices), coils (random coils) and rods (extended strand) along the primary sequence, and where (---) indicate sequence interruptions.

Three dimensional homology models of protein structures were constructed using the Swiss Model software\(^{(18, 19)}\) with templates obtained from Protein Databank\(^{(20)}\). Homology modeling relies on the identification of protein structures likely to resemble the structure of a protein in question, and on the production of an alignment that maps residues in the protein of known three-dimensional structure to residues in protein in question.

**RESULTS AND DISCUSSION**

Relationships between proteins may be considered into two different ways. Proteins may display the same activity and show sequence similarity suggesting a common ancestral origin. Inferential statistics are frequently used to evaluate their relatedness. When any two proteins sequences or domains within proteins display statistically significant similarities, the proteins are considered as being homologous.

Proteins may also display the same activity but lack sufficient similarity to imply common origin. Such proteins are said to have analogous activity. The implication is that analogous proteins followed
evolutionary pathways from different origins to converge upon the same activity. Thus, analogous proteins are considered a product of convergent evolution. That is, analogous proteins have homologous activity but heterologous origins. Evidence that two proteins are related by analogy frequently involves comparisons of protein secondary and tertiary structure, as well as comparison of primary sequences structure, which may not show statistically insignificant relationships. In this study, evidence of possible protein analogies was considered.

Analysis of the human genome revealed that the enzyme glyceraldehyde-3-phosphate dehydrogenase (EC 1.2.1.12) resembles the CysP periplasmic sulfate-binding protein in bacteria. The GAPDH protein for glyceraldehyde-3-phosphate dehydrogenase is 335 residues in length. GAPDH is a key enzyme in glycolysis that catalyzes the first step of the pathway by converting D-glyceraldehyde 3-phosphate (G3P) into 3-phospho-D-glyceroyl phosphate\(^\text{21}\). The enzyme also participates in nuclear events\(^\text{22}\).

Glyceraldehyde-3-phosphate dehydrogenase from Homo sapiens was found to display 49.6% similarity to CysP of Escherichia coli with 13% identical residues and where 36% constituted conservative or semiconservative sequence matches. To achieve optimal similarity between the proteins, 7 small gaps were introduced in the alignment (see Appendix). A reliability estimate, derived from a set of 1300 sequences, and belonging to 58 proteins produced a probability of 99.6% for the likelihood that the two proteins belongs to the same family of the proteins\(^\text{16}\). The amino acid sequences for glyceraldehyde-3-phosphate dehydrogenases are strongly conserved across all phyla\(^\text{23}\). Thus, results found in Homo sapiens may extend to other species.

Secondary structural comparisons indicate the alpha helical characteristics of the CysP protein is 41.4% and the alpha helical characteristics of glyceraldehyde-3-phosphate dehydrogenase is lower than that observed in CysP protein consisting of 21.8% of the protein. The enzyme has a higher extended strand potential comprising 26.8% of the protein as compared to CysP with 13.9%. Both proteins have the same random coil potential. In GAPDH, the random coil potential was 48.9% as compared to CysP with 43.5%. The overall distribution of structural characteristics was found to be similar through much of the two proteins. Residues 110-240 of GAPDH are highly similar to residues 100-230 of the periplasmic binding protein (Figure 1). Residues 240-259 of the dehydrogenase enzyme have the same distribution coil and α-helical characteristics as 240-260 of the periplasmic binding protein. Both proteins lack regions predicted to contain beta-pleated sheet structures.

Figure 1. Homology models of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and CysP periplasmic binding proteins. The Venus Flytrap motif is indicated and the proposed motif in GAPDH is shown.

To support the observations derived from predictions of secondary structure, atomic-resolution models of the proteins were determined by homology modeling studies. It has been shown that protein structures are more conserved than protein sequences amongst homologues, but
sequences falling below a 20% sequence identity can have very different structure \cite{24}.

Homology models of the three dimensional structure of both proteins has been determined where the structures of 1sbp and 1u8f (Protein Data Bank accession numbers) \cite{20} were used as templates for CysP and GAPDH, respectively. As shown in Figure 1, the structure of CysP has a bilobate region and a hinge region that is characters of all proteins that adapts to the Venus Flytrap motif. GAPDH also appears to have a structure similar to the general structure of a Venus Flytrap motif. It is suggested that, owing to the bilobate region and a hinge region, GAPDH has a fold structure characteristic of periplasmic binding proteins.

The results of this study support the idea that sequence and structural analogies exist between glyceraldehyde-3-phosphate dehydrogenase and periplasmic binding proteins. A Venus Flytrap motif in the structure of glyceraldehyde-3-phosphate dehydrogenase might support interaction of glyceraldehyde-3-phosphate with the enzyme. Further studies are required in order to determine the role of this structure in enzyme function.

**ACKNOWLEDGEMENTS**

The author wishes to thank the anonymous referee and Dr. Edward Smith for many helpful suggestions and comments on the preparation of this manuscript.
APPENDIX

Comparative sequence analysis and predicted structural comparison of Glyceraldehyde-3-phosphate dehydrogenase (GADPH) and the CysP periplasmic binding protein. Identical residues are indicated by (*). Sites where amino acid residues are related by conservative replacements are indicated by (:). Sites were related by semi-conservative replacements are indicated by (.). Replacement groups for conservative replacements (in one-letter code) are (STA, NEQK, NHQK, NDEQ, QHRK, MILY, MILF, HY, FYW) and for semi-conservative replacements are (CSA, ATV, SAG, STNK, STPA, SNDEQK, NDEQHE, NEQHRK, FYLM, HFY). Secondary structural elements are indicated by solid color coils (α-helices), coils (random coils) and rods (extended strand) along the primary sequence. Sequence interruptions are indicated by dashed line (---).
REFERENCES


