Journal of the South Carolina Academy of Science 2(1):1-17 Fall 2004

BIOINFORMATICS: AN UNDERGRADUATE RESEARCH/TEACHING TOOL

Abhishek Agrawal[1] and Valgene L. Dunham[2]

[1] Keck Graduate Institute, 535 Watson Drive, Claremont, CA 91711
[2] Department of Biology, Coastal Carolina University, Conway, SC 29528
dunham@coastal.edu

ABSTRACT

The science of bioinformatics can be effectively used in undergraduate teaching and research related to biology, biochemistry and computer science. This paper employed the biosynthetic pathway leading to the synthesis of the amino acid lysine to illustrate the software and information available on the internet, concerning protein and gene sequences. The enzymes of the pathway were identified and compared in three organisms: the bacterium Escherichia coli, Saccharomyces cerevisiae (yeast) and the plant Arabidopsis thaliana. The amino acid and nucleotide sequences of the key regulatory enzymes aspartate kinase and homoserine dehydrogenase found in these organisms were employed to illustrate the use of readily available software. The sequence of eight different genes for the two enzymes in the three organisms was identified including two different bifunctional proteins with both activities in both E. coli and Arabidopsis. Conservation of amino acid sequences coded for by these eight genes in these diverse organisms was indicated using sequence alignment software. Possibilities for the use of bioinformatics in instruction are illustrated including obtaining motif and domain information, construction of phylograms and predicting the structure of proteins based on included data bases and software. Internet locations for software employed are included in the appendix.

INTRODUCTION

The rapidly expanding amount of information concerning genomes, protein structure and biochemical network maps in a number of organisms has resulted in the interdisciplinary science of bioinformatics. Public access on the internet to genomes, biological data banks and the software to analyze the information may be used as learning tools in numerous undergraduate courses both in biology and computer science. This paper presents a model for how enzyme sequence data available on the internet can be used by undergraduates to study metabolic pathways, the interrelationships between metabolic pathways over ranges of different organisms, and the comparative characteristics of specific enzymes within these pathways.

The metabolism of aspartic acid is an extremely important metabolic pathway that is necessary for the production of other amino acids (lysine, threonine, methionine, isoleucine and leucine; depending on the pathways in specific organisms). Interest in this pathway by the senior author of this paper was initiated by physiological studies on the effects of this family of amino acids on the growth and development of the liverwort Marchantia polymorpha (Dunham and Bryan, 1969, 1971). More than 30 years later, significant information has been accumulated in a diversity of organisms related to the enzymes involved and the diversity of genes required for their synthesis and regulation (Cohen and Saint-Girons, 1987; Galili, 2002). This pathway exhibits numerous methods of control including specific feedback inhibition by end products of the pathway, multiple forms of these regulated enzymes that have varying sensitivity to feedback

inhibitors (Bryan, 1990), multifunctional enzymes (Muehlbauer et al., 1994) and different pathway intermediates depending on the organism. In addition, gene expression of a key enzyme in the pathway in plants is subject to modulation during growth and development within specific tissues (Zhu-Shimoni et al., 1997).

This pathway was selected as a model for the use of the techniques of bioinformatics at the undergraduate level because of its diversity in regulation, known sequences of genes coding for enzymes in the pathway from several, diverse organisms and its integration with other essential metabolic pathways.

MATERIALS AND METHODS

The selection of the appropriate software programs to employ depends on the intended instructional/research objectives. In this paper in which these tools are to be illustrated, KEGG (Kyoto Encyclopedia of Genes and Genomes) was employed as the common metabolic database. KEGG compiles a reference network map using specific organism databases like EcoCyc, AraCyc and publications in biochemistry. Since KEGG is a secondary source of information, specific databases were used in this paper to verify information. KEGG is linked to other database retrieval systems such as DBGET, LinkDB and GenomeNet. These are all flat-file data systems but also include GIF images of KEGG pathways, Java graphics and 3D images of protein structures. In addition, these systems are also linked to sequence interpretation tools such as BLAST, FASTA, MOTIF, CLUSTALW and Pfam. Pfam, a database of protein families assembled using multiple sequence alignments, is primarily used to identify and assign a function to proteins and is often the first step in homology modeling. Pfam is often used to identify probably domains in a known protein sequence.

To illustrate instructional use of bioinformatics in the metabolism of lysine, KEGG was employed using the following steps:

1. From the KEGG website (see website appendix), Protein Network was selected
2. Under 1.5 Amino acid biosynthesis, lysine biosynthesis was selected.

The lysine biosynthesis Reference Pathway (see appendix) provides numerous possibilities for instruction/research. The pathway indicates the metabolites involved in the pathway, enzymes included in sequence listed by their EC classification, and associated metabolic pathways. Selection of any of the compounds indicates the molecular structure, name and organic nomenclature. Selection of an enzyme results in the enzyme's reaction, systematic name and class, a list of substrates and products, references, and for purposes of this paper, a list of genes from different organisms that when selected, will display the amino acid sequence of the enzyme and the nucleotide sequence of the gene from that organism.

Selection of the Ortholog Table on the Reference Pathway page results in a list of organisms with the enzymes of the pathway. Under each organism listed are the letters P, G and T. Selection of P results in a Reference Pathway outlined for that specific organism. G results in the known genetic map of the organism as the number of nucleotides, protein genes and RNA genes as well as a genome map browser. The map browser allows an area of a chromosome to be selected resulting in the nucleotide sequence and the sequence of amino acids of the encoded protein. In addition, selection of MOTIF indicates any motifs present in the sequence and will allow a search for that motif in other genes. T results in a list of the pathway enzymes in that organism with their amino acid sequence. To compare amino acid sequences of enzymes from different organisms, sequence alignment software was employed (CLUSTRAL W-see website index). Functional domains of proteins within the lysine pathway were determined and

compared using Superfamily, a HMM profile library built using the hierarchy of protein structures established by Structural Classification of Proteins (SCOP-see appendix).

RESULTS

To illustrate the use of the tools of bioinformatics in teaching biology, biochemistry and bioinformatics courses at the undergraduate level, KEGG was employed to investigate the metabolic sequences of lysine biosynthesis in a diverse group of organisms: *Escherichia coli* K12, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*. These organisms were selected on the basis of known genome sequences and the involvement of regulatory enzymes in the biosynthesis of lysine. The KEGG reference page for lysine biosynthesis includes the amino acid and gene sequence for six enzymes present in all three organisms (Fig. 1). Due to the involvement of the first three enzymes in the regulation of the pathway, they were selected for use in illustrating the software tools: aspartokinase (2.7.2.4), aspartic semialdehyde dehydrogenase (1.2.1.11) and homoserine dehydrogenase (1.1.1.3). The other three enzymes present in all three organisms are of general interest and are transferases and the aminoacyl-tRNA synthetase that "charges" lysyl-tRNA (Fig. 1).

From the lysine pathway chart (Fig. 1), the amino acid sequence of the protein and nucleotide sequence of the gene that codes for each of the three enzymes may be obtained by selecting the enzyme's EC number.

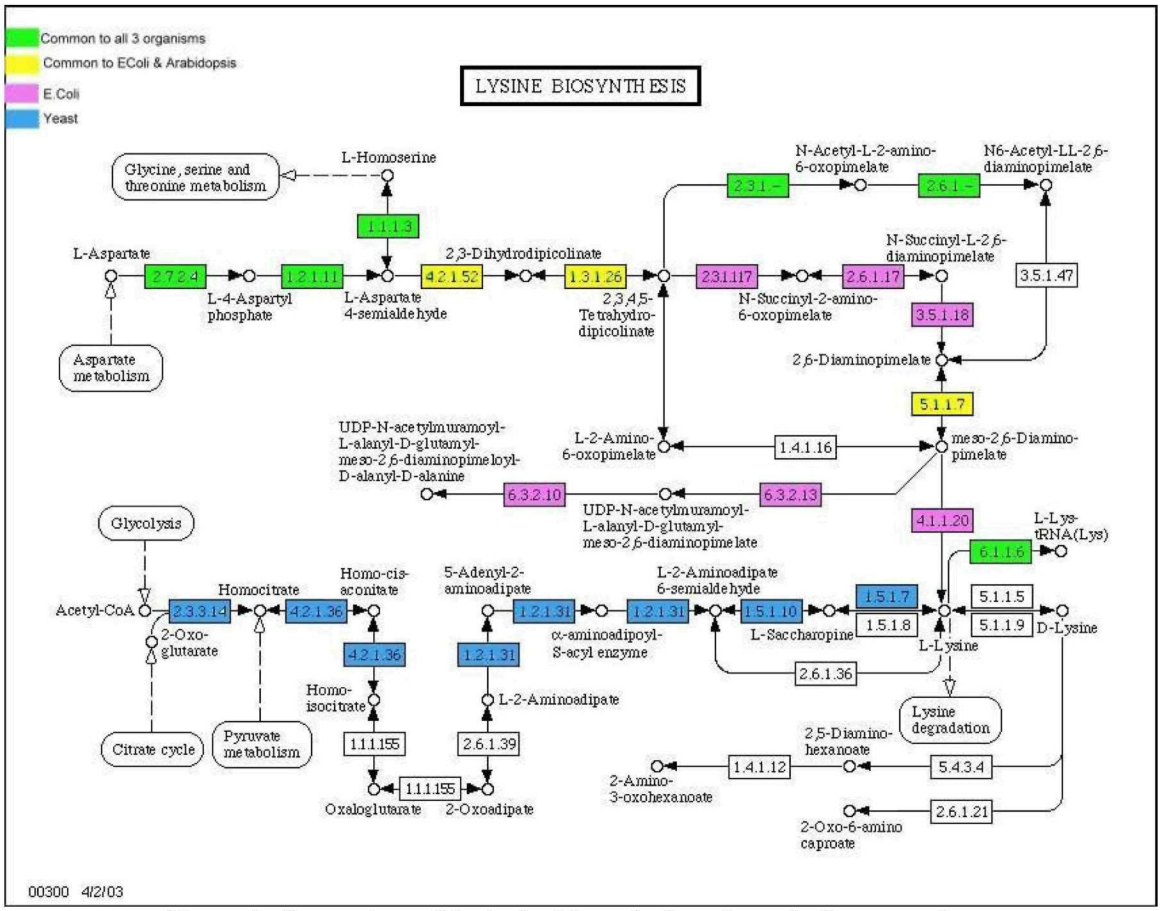


Figure 1. Comparison of the lysine biosynthetic pathway in three organisms.

Upon investigation of the total number of genes in the three organisms related to the three selected enzymes at the beginning of the pathway, several forms of aspartate kinase are present in the three organisms. As indicated in Table 1, both *E. coli* and Arabidopsis each contain two distinct genes that code for a bifunctional protein with both aspartate kinase and homoserine dehydrogenase activities. In addition, *Arabidopsis* also has two genes for a monofunctional aspartate kinase whereas *E. coli* and yeast each have only one such gene.

Table 1. Multiple genes for initial enzymes in the lysine biosynthetic pathway.

Enzyme	Organism	Gene	A. acids	Nucleotides
Bi-functional protein: Aspartate kinase – Homoserine dehydrogenase	*E. coli*	b3940	810	2433
		b0002	820	2463
	Arabidopsis	at1g31230	911	4978
		at4g19710	916	4817
Aspartate kinase (monofunctional)	*E. coli*	b4024	449	1350
	Yeast	YER052C	527	1584
	Arabidopsis	at5g13280	569	3399
		at5g14060	544	3793
Aspartate semialdehyde dehydrogenase	*E. coli*	b3433	367	1104
	Yeast	YDR158w	356	1098
	Arabidopsis	at1g14810	375	2266

To illustrate software utilization in the study of these enzymes and possible differences in the three organisms studied, aspartate kinase and homoserine dehydrogenase were selected for amino acid sequence comparisons. This selection was based on the presence of both mono and bifunctional proteins in both *E. coli* and *Arabidopsis*. To compare amino acid sequence of the proteins or the nucleotide sequence of the genes, eight sequences were compared using sequence alignment software CLUSTAL W (see appendix). These eight sequences include the four genes from Arabidopsis (2 bifunctional proteins and two monofunctional aspartate kinases activities), 3 genes from E. coli (2 bifunctional proteins and one monofunctional aspartate kinase) and the monofunctional aspartate kinase from yeast. Several regions of these eight genes were of interest with respect to coding for identical or nearly identical amino acid sequences (Fig. 2).

The two monofunctional aspartate kinase genes in the plant are located on chromosome 5 whereas the genes for the bifunctional proteins are located on different chromosomes, chromosome 1 and 4 (Fig. 2). A more detailed analysis may be made by limiting the analysis of sequence alignment to either the bifunctional or monfunctional enzymes. To illustrate the probability of increased homology, just the bifunctional aspartate kinase enzymes in *E. coli* and *Arabidopsis* were subjected to sequence alignment (Fig. 3).

The proteins from *E. coli* and *Arabidopsis* have several regions that are highly conserved, especially the region indicated in Fig. 3. The amino acid sequences in the identical regions and other areas of conservation may be assessed for possible motif and domain significance and function. For example, MOTIF software locates the aspartate kinase domain of the bifunctional enzyme at the N terminus of the protein and the homoserine dehydrogenase domain toward the C terminus which contains the conserved sequences in Fig. 3. MOTIF also locates a NAD binding site in the plant genes (amino acids 566-701 in at1g31230 and amino acids 571-706 in at4g19710).

```
At5g13280       GAVTTLGRGGSDLTATTIGKALGLKEIQVWKDVDGVLTCDPTIYKRATPVPYLTFDEAAE
341
At5g14060       CAITTLGRGGSDLTATTIGKALGLREIQVWKDVDGVLTCDPNIYPGAQSVPYLTFDEAAE
337
b4024           GRTTTLGRGGSDYTAALLAEALHASRVDIWTDVPGIYTTDPRVVSAAKRIDEIAFAEAAE
250
At1g31230       NIPTTLKRDGSDFSAAIMSALFRSHQLTIWTDVDGVYSADPRKVSEAVVLKTLSYQEAWE
338
At4g19710       NIPTTLKRDGSDFSAAIMGALLRARQVTIWTDVDGVYSADPRKVNEAVILQTLSYQEAWE
343
b3940           GETVLLGRNGSDYSATQIGALAGVSRVTIWSDVAGVYSADPRKVKDACLLPLLRLDEASE
259
b0002           GELVVLGRNGSDYSAAVLAACLRADCCEIWTDVDGVYTCDPRQVPDARLLKSMSYQEAME
256
Yer052c         GLLNGVGRGYTDLCAALIAVAVNADELQVWKEVDGIFTADPRKVPEARLLDSVTPEEASE
282
                     : *. :*  *: :.         :*.:* *: : **     *  :  :   ** *
```

* = high probability of sequence homology or identical

: = less probability of sequence homology, conserved substitutions have been observed (refer to color codes)

. = least probability of sequence homology (semi-conserved)

At5g13280 = single function protein Arabidopsis
At5g14060 = single function protein Arabidopsis
b4024 = single function protein E. coli
At1g31230 = bifunctional protein Arabidopsis
At4g19710 = bifunctional protein Arabidopsis
b3940 = bifunctional protein E. coli
b0002 = bifunctional protein E. coli
Yer052c = single function protein yeast

AVFPMILW	RED	Small (small+ hydrophobic (incl.aromatic -Y))
DE	BLUE	Acidic
RHK	MAGENTA	Basic
STYHCNGQ	GREEN	Hydroxyl + Amine + Basic - Q
Others	Grey	

Figure 2. Section of the amino acid sequence alignment of Aspartate kinase from three organisms.

```
At1g31230       YYDWLLRGIHVVTPNKKANSGPLDQYLKIRDLQRKSYTHYFYEATVGAGLPIISTLRGLL
712
At4g19710       YYDWLRKGIHVITPNKKANSGPLDQYLKLRDLQRKSYTHYFYEATVGAGLPIISTLRGLL
717
b0002           YADFLREGFHVVTPNKKANTSSMDYYHQLRYAAEKSRRKFLYDTNVGAGLPVIENLQNLL
623
b3940           YLDFASHGFHVISANKLAGASDSNKYRQIHDAFEKTGRHWLYNATVGAGLPINHTVRDLI
617
                * *:  .*:**::.** *.:.  : * :::   .*:  :::*::.******:  .::.*:

At1g31230       ETGDKILRIEGIFSGTLSYLFNNFVGTRSFSEVVAEAKQAGFTEPDPRDDLSGTDVARKV
772
At4g19710       ETGDKILRIEGICSGTLSYLFNNFVGDRSFSEVVTEAKNAGFTEPDPRDDLSGTDVARKV
777
b0002           NAGDELMKFSGILSGSLSYIFGKLDEGMSFSEATTLAREMGYTEPDPRDDLSGMDVARKL
683
b3940           DSGDTILSISGIFSGTLSWLFLQFDGSVPFTELVDQAWQQGLTEPDPRDDLSGKDVMRKL
677
                ::** :: :.** **:**::* ::    .*:* .  * : * ********** ** **:
```

Figure 3. Amino acid sequence alignment from the most conserved region of the bifunctional aspartate kinase and homoserine dehydrogenase.

To further examine the relationships between the sequences of these two enzyme activities in the three organisms, the amino acid sequence alignments can be employed by CLUSTAL W software to generate phylogenic trees. Although more amino acid sequences need to be added to produce a more comprehensive comparison, Fig. 4 illustrates sequence relationships with just three organism and eight genes.

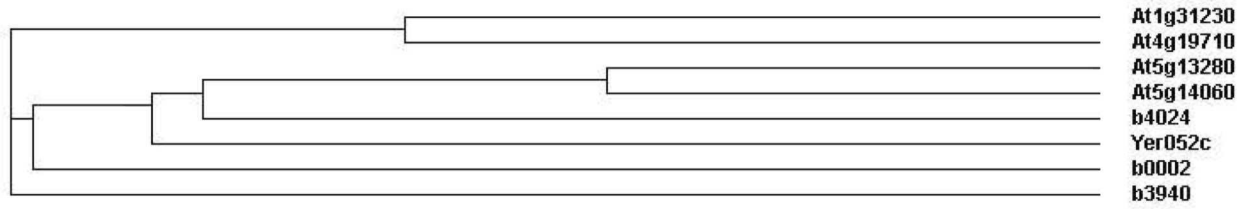


Figure 4. Phylogram of genes coding for aspartate kinase and homoserine dehydrogenase in three organisms.

As expected, similarities in sequence place the two plant bifunctional enzymes together (at1g31230 and at4g19710) and the two monofunctional plant genes for aspartate kinase with the monofunctional genes for the enzyme in yeast and bacteria. It is of interest that this simple phylogram places the E. coli b0002 bifunctional enzyme closer in sequence to the monofunctional enzymes in yeast and in the plant than to the other biofunctional enzyme in E. coli (Fig. 4). This result may be due to gene size differences in which a smaller paralog will have a lower similarity or the possibility of lateral gene transfer during the evolution of the genes.

An investigation of the functional domains of the aspartate kinase and homoserine dehydrogenase activities in these organisms was made using Superfamily software. This software compares domains within the amino acid sequence of the proteins with a database of known functional domains in other proteins and illustrates the repeated use of these structures in numerous and diverse proteins. More importantly, these domain studies may illustrate that a functional domain shared by proteins may not share similar or identical amino acid sequences.

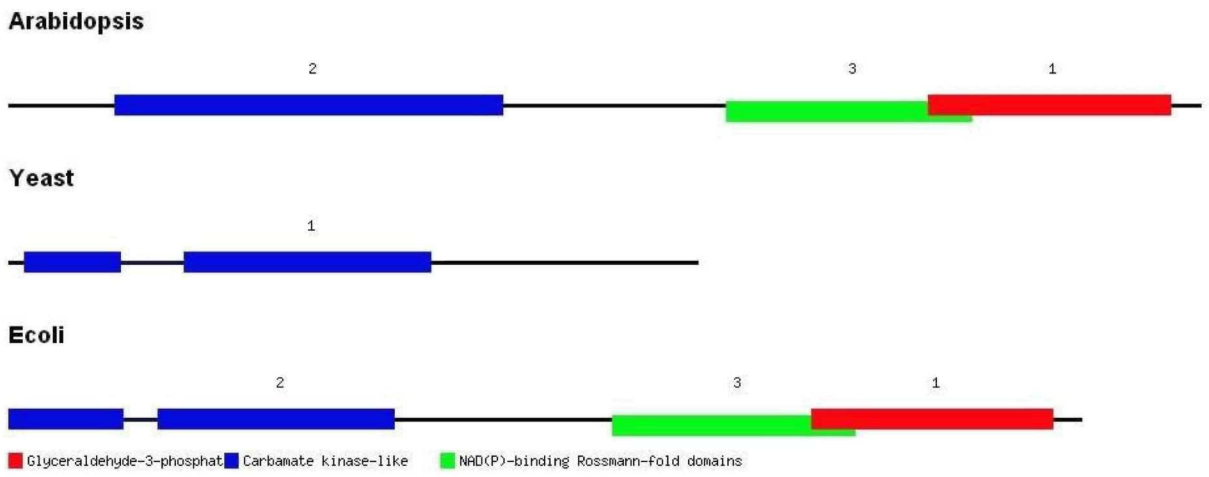
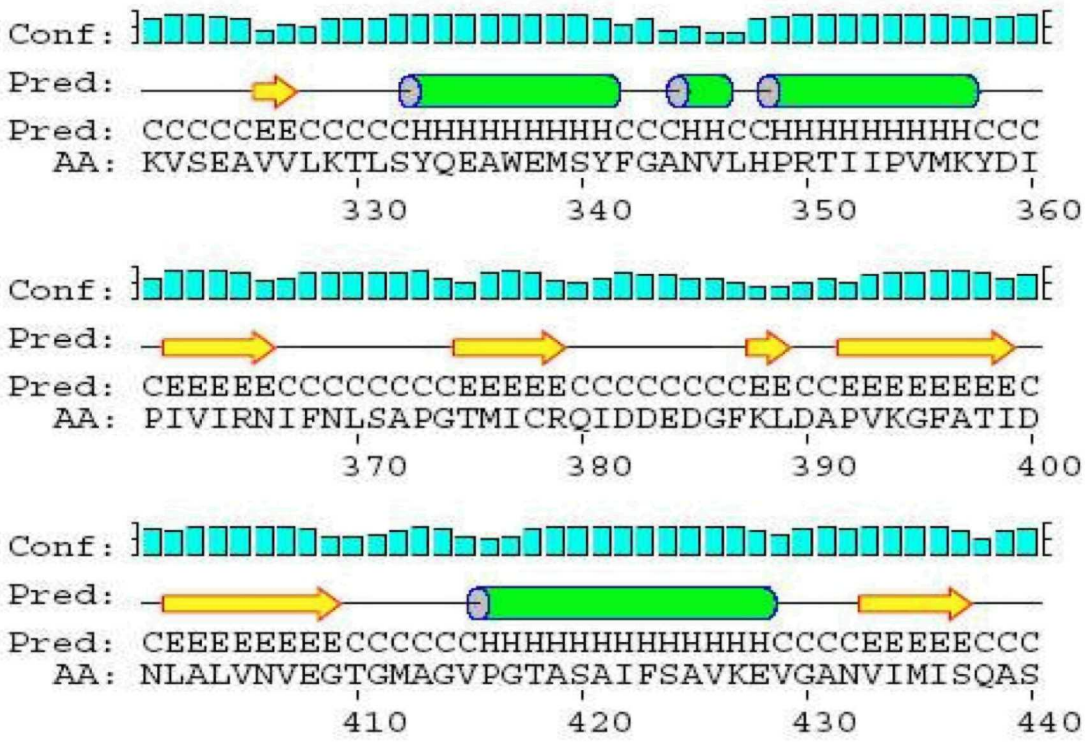


Figure 5. Domain comparisons in aspartate kinase and homoserine dehydrogenase with a domain database. Arabidopsis and E. coli enzymes are bifunctional proteins.

For example, the binding site for NAD(P) (Number 3 in Fig.5) illustrates the presence of a general domain involving the Rossmann fold which need not have any amino acid similarities (Eventoff and Rossmann, 1975). A dehydrogenase domain (red) and a kinase domain (blue) are also indicated in this comparison (Fig. 5).

The use of sequence comparison software is especially informative when an obtained protein sequence (from either research or for instructional purposes) is used to determine its possible secondary structure. This information, coupled with domain software illustrated above, may be used in determining the overall probable three dimensional structure and function of the protein. Superfamily and Pfam were used for homology modeling, i.e. Figure 5. For secondary structure, software available on the protein structure prediction server (PSIPRED) was used (see appendix). A possible secondary structure of the bifunctional aspartate kinase-homoserine dehydrogenase of Arabidopsis (at1g31230) was obtained (Fig. 6). This software makes predictions of secondary structure based on the three dimensional structure of known proteins. A small section of the enzyme that illustrates both α-helices and β-sheets was selected from the dehydrogenase domain (characterized by β-sheets on one side of the molecule and α-helices and random coils on the other).



Conf: indication of confidence
Pred: yellow arrows = β-sheet; green cylinders = α-helix
Pred: H = α-helix; E = β-sheet; C = intermittent sequences
AA : amino acid sequence

Figure 6. Possible secondary structure in a section of the bifunctional enzyme from Arabidopsis.

DISCUSSION

The use of the software and information presented in this paper shows how bioinformatics may be employed at the undergraduate level. Information presented concerning the enzymes involved in lysine biosynthesis represents only a small section of the software and sequences cataloged by and linked to KEGG. For example, in addition to the section on protein networks, other KEGG sections include Genetic Information Processing (transcription, translation, replication and repair), Environmental Information Processing (membrane transport, signal transduction, immune system), Cellular Processing (motility, growth, apoptosis, behavior) and a section on Human Diseases (neurodegenerative). Each one of these areas has topics for further investigation similar to the protein network used in this paper. These include gene databases which allow for investigation of chromosome location, sequence matching and protein domains. Ligand databases include information on the structure of compounds, glycans, reactions and molecular classification.

Based on the information presented in this paper, the software and genomic information available might be effectively used in undergraduate courses as follows:

Undergraduate course	**Possible use**
General Introductory Biology	Introduction and review of basic information: Amino acid structure Protein structure Enzyme function, basic kinetics Relationship between DNA nucleotide sequence and amino acid sequence of proteins Metabolic pathways and regulation Bioinformatics
Upper Level Biology Courses: (Cell, Molecular Biology, Biochemistry, Genetics)	Review of introduced concepts (above) Phylograms Sequence alignment comparisons Protein motif and domain studies Predicting protein structure Molecular evolution
Introductory Bioinformatics	Use of software Application of AI concepts – Machine Learning, pattern recognition, clustering and classification.

The use of the software and available sequences in research assignments related to undergraduate courses and undergraduate research may be expanded to include an analysis of new protein sequences or sequences not included in the software employed in this paper. For example, two bifunctional enzymes (aspartate kinase-homoserine dehydrogenase) have also been characterized in other plants including maize (Muehlbauer et al., 1994). Undergraduate research and bioinformatics students could compare the sequences and structures of the maize enzymes with respect to the NADPH binding domains (amino acids 568-573), a KFGG region (amino acids 98-101) near the N-terminus of aspartate kinase domains and an interface region (amino acids 342-566 and 339-563, respectively; Muehlbauer et al., 1994) with the bifunctional enzymes I and II in Arabidopsis.

The application of bioinformatics to the undergraduate curriculum in biology and computer science allows for numerous links between students learning to create software and students who must use the software in studying molecular biology, genetics and molecular evolution. An understanding of the creation and use of the software is becoming an essential step in the education of the modern biologist.

ACKNOWLEDGEMENTS

This paper is in appreciation for the excellent mentoring the senior author received while he was the first Ph.D student of Dr. Jack Bryan, Department of Biology, Syracuse University. The authors also acknowledge helpful comments by Dr. Benjamin Matthews, USDA-ARS-PSI. Dr. Matthews also completed his Ph.D research under the direction of Dr. Bryan.

LITERATURE CITED

Bryan, JK. 1990. Advances in the biochemistry of amino acid biosynthesis. In: Miflin, BJ, Lea, PJ, editors. The Biochemistry of Plants, Vol 16. New York. Academic Press, p. 161-195.

Cohen, GN, Saint-Girons, I. 1987. Biosynthesis of threonine, lysine and methionine. In: Neidhardt, FC, editor. Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology. Washington, D.C.: American Society of Microbiology. p. 429-444.

Dunham, VL, Bryan, JK. 1969. Synergistic effects of metabolically related amino acids on the growth of a multicellular plant. Plant Physiology 44: 1601-1608.

Dunham, VL, Bryan, JK. 1971. Synergistic effects of metabolically related amino acids on the growth of a multicellular plant. II. Studies of 14C-amino acid incorporation. Plant Physiology 47: 91-97.

Eventoff, W, Rossmann, MG. 1975. The evolution of dehydrogenases and kinases. Critical Review of Biochemistry 3: 111-140.

Galili, G. 2002. New insights into the regulation and functional significance of lysine metabolism in plants. Annual Review of Plant Biology 53: 27-43.

Muehlbauer, GJ, Somers, DA, Matthews, BF, Gengenbach, BG. 1994. Molecular genetics of the maize (Zea mays L.) aspartate kinase-homoserine dehydrogenase gene family. Plant Physiology 106: 1303-1312.

Zhu-Shimoni, XJ, Lev-Yadun, S, Matthews, BF, Galili, G. 1997. Expression of an aspartate kinase homoserine dehydrogenase gene is subject to specific spatial and temporal regulation in vegetative tissues, flowers and developing seeds. Plant Physiology 113: 695-706.

APPENDIX

Query → Biological Database → Tools to analyze the information.

1. Query: A query could be defined as a low level abstraction of a biologically relevant question, which could be translated into searchable keyword. For instance searching for a gene / protein / pathway / protein or gene function, etc.
2. Types of biological databases.
    a. Sequence Databases
    b. Structural Databases: The key to the functionality of a Bio-molecule lies in its function. The field of structural bioinformatics aims to profile the structure of all bio-molecules using physical and computational methods.
    c. Network Databases
        i. Metabolic Network Databases
            1. Protein Interaction Databases
            2. Gene Interaction Databases
        ii. Signal Transduction Databases
    d. Topical Databases- metabolic networks, genes and diseases etc.
3. Types of analytical tools: To analyze the growing amount of biological data, computational tools have been designed using concepts from various science and engineering fields. For instance, data clustering techniques for grouping similar objects had been developed for clustering celestial objects, now the same principles are being used to identify and group similar genes, proteins and organisms. Similarly, machine learning techniques like Hidden Markov Models which were used in speech recognition are now being used for protein homology prediction.

4. Tools available freely over the internet.
    a. Search
    b. Sequence Alignment
    c. Structure prediction
    d. Clustering: is used to group similar entities together using some of their characteristics. For instance, sequence similarity is used to cluster homologous proteins / genes.
    e. Visualization Tools
    f. Biochemical network modeling

ILLUSTRATION OF USE IN THIS PAPER

1. Query: Evolutionary conservation in the lysine biosynthesis pathway of Arabidopsis, E. coli and yeast.
2. Biochemical network maps: KEGG lysine biosynthesis pathway reference map was used to identify and evaluate the common proteins.
3. Enzyme Commission nomenclature.
4. Amino Acid Sequence files for Enzyme
    a. 2.7.2.4
    b. 1.2.1.11
    c. 1.1.1.3
5. Multiple Sequence Alignment & phylogenetic tree
    a. 2.7.2.4
    b. 1.2.1.11
    c. 1.1.1.3
6. Protein Structure prediction
    Predicting the protein structure computationally is not feasible due to the exponential number of possibilities based on the interaction of various amino acids i.e. finding the most stable energy state of an exponential number of possibilities. As a result, the problem can be broken down into the identification of the functional elements of the proteins:
        1. Homology based function prediction
            1. Sequence Homology
            2. Structural Homology
        2. Identification of Domains: Using homology, the probable function of a protein can be predicted by identifying the evolutionarily conserved functional units.
        3. Predicting the Secondary Structure
        4. Transmembrane topology
        5. Fold prediction

Homology – Domain Identification (structural)

2.7.2.4 Results: 1 , 2

1.2.1.11

1.1.1.3

The results of homology analysis identified common domains in the group of proteins selected.

Resource: Super-family[1], Pfam[2], HMM

Secondary Structure

ii. Bi-functional Enzyme: Aspartate Kinase – Homoserine Dehydrogenase in Arabidopsis: At1g31230 Results –[TXT] [IMAGE]

Resource: Protein Structure Prediction Server

WEBSITES

Types of biological databases.

Single instances of each type database are listed here, for a complete list please check the note.

Sequence Databases

Nucleotide - GENEBANK

Protein Sequences – Protein Information Resource (PIR)

Structural Databases

Protein – Protein Data Bank (PDB)

Carbohydrates - CARBBANK

Metabolic and Signaling Databases

1. Pathway and genome database - BioCyc Database
2. Database profiling binary relationship between proteins and other bio-molecules - Protein Interaction Database

Topical Databases

Genes and Diseases –HyperGenome Human Genome and Disease database.

Micro-array and gene expression databases - ArrayExpress

Enzymatic information: Databases listing enzyme kinetics, isolation procedure, chemical reactions. - BRENDA

Scientific Literature reference database - PubMed

Metabolic Network Models - KEGG

Note: Complete categorized and updated list of all Biological databases can be found on the webpage maintained by the Oxford Journal's online database resource page.

Website: http://www3.oup.co.uk/nar/database/subcat/1/2/

Tools

Tools available for the given categories.

Search

Blast

UniProt

Sequence Alignment

ClustalW

Tcoffee

Structure prediction

PSIPRED

DALI

Visualization Tools

Protein Explorer

Deep View

Biochemical network modeling

JDesigner

Cytoscape

Research Hub resources

NCBI: National Center for Biotechnology Information

The NCBI site has two of the most heavily used resources in Bioinformatics – BLAST: Fastest homology searching tool and PUBMED. NCBI also maintains an integrated web based database retrieval system – ENTRENZ: The Life science search engine, which is equivalent to Google of biological data.

Website: http://www.ncbi.nlm.nih.gov/Entrez/index.html

NCBI has a bookshelf of the best books in Molecular Biology.

Website: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books

EBI: European Bioinformatics Institute, established by EMBL – European Molecular Biology Laboratory.

The site classifies the various tools in a visual manner and enables the user to have a complete view of all the Bioinformatics tools. This facilitates a novice user to put usage requirements in perspective to all the tools available and provide scheme for tool usage progression. Moreover, EBI has developed a learning resource to supplement all the tools called –2can

Website: http://www.ebi.ac.uk/services/index.html

GenomeNet: Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Some of the most sort after resources available here are: KEGG, Ligand Database, MOTIF, Gene Catalog.

Website: http://www.genome.ad.jp/

BiBiServe: Bielefeld University Bioinformatics Server

Website: http://bibiserv.techfak.uni-bielefeld.de/

Biological Initiatives

Each of the following initiatives provided an impetus to the growth of biological knowledge.

The Human Genome Project

Website: http://www.nhgri.nih.gov/

The Human Brain Project: Provides a central source for various brain models being developed and aims to integrate all of them to reveal the architecture and functioning of the human brain. The efforts of the human brain project have been aptly labeled as – Neuroinformatics.

Website: http://www.nimh.nih.gov/neuroinformatics/index.cfm

Genomes for Life: funded by the US department of energy is focusing in the area of systems biology.
Website: http://doegenomestolife.org/

LEARNING RESOURCES

2can: Bioinformatics Educational Resource.
The European Bioinformatics Institute has compiled a set of tutorials covering all the major tools available on its site. The tutorials provide a brief overview of the algorithm being used, along with complete workflows (images) on how to use the tools effectively. Since, the site is maintained and developed by EBI, most of the information on it is accurate and is being constantly updated.
Website: http://www.ebi.ac.uk/2can/home.html

S*Star.org Alliance: Is a global alliance of several universities, contributing several lectures on Bioinformatics from researchers around the world. Therefore, this site is not a quick – howto. The lectures provide an overview of a specific area like – comparative genomics. Protein-ligand modeling or Dynamic programming used in HapMap project and they ought to be supplemented with external reading.
Website: http://www.s-star.org/

Bioinformatics Lectures: This resource is primarily for students who are computationally and mathematically oriented.
Website: http://lectures.molgen.mpg.de/online_lectures.html

Nomenclature

Efforts are under way to provide a controlled and universal vocabulary for the enormous amount of biological data being generated.

Gene Ontology (GO): The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.
Website: http://www.geneontology.org/

Enzyme Commission (EC)
Website: http://www.chem.qmw.ac.uk/iubmb/enzyme/

RESOURCE DESCRIPTION

This section contains a brief description of all the resources used in the illustrative example. The description covers the concept being used by the tool, its features and references for further exploration.

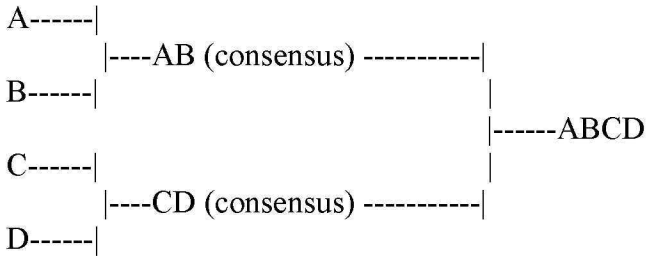
1. CLUSTALW
2. FLAT-FILE
3. HMM
4. KEGG
5. MOTIF
6. PFAM
7. Protein Structure Prediction Server
8. SCOP

9. SUPERFAMILY

CLUSTALW

CLUSTALW is the latest in the Clustal series of tools for performing multiple sequence alignments developed by Dr. Des Higgins and collaborators. Basically, there are four criterions for producing an alignment: structural, evolutionary, functional and sequence similarity. Clustal uses sequence similarity as it is easiest to implement computationally. It uses progressive sequence alignment to produce a multiple sequence alignment. The Phylogram produced by Clustal can be used to step through the algorithm used by Clustal.

Suppose there are 4 sequences: A, B, C, D. Initially a pairwise sequence alignment score is generate for each of ${}^4C_2$ possibilities and the pairs with highest scores are grouped. This step is essentially followed iteratively till all the sequences get aligned.



Multiple-alignment can be used to identify patterns like domains, motif and binding sites or phylogeny or protein family classification, etc.

Website: http://www.ebi.ac.uk/clustalw/

FLAT-FILE

A flat-file is text file mirroring the database structure with appropriate tags. So that programs can interpret the data. A HTML file has tagged information but it is not organized. A flat-file has structured information with appropriate tags. It is hard to update / insert / delete information from these files.

```
gene                 1..2076
                    /gene="ttll1"
                    /note="synonym: MGC63993"
                    /db_xref="LocusID:393994"
                    /db_xref="ZFIN:ZDB-GENE-040426-1326"
```

Tags: /gene, /note,

Information for reader: gene 1...2076

HMM: Hidden Markov Model

Basically HMM tools identify the statistical pattern in a group of unaligned sequences, and build a profile of those sequences. This HMM profile of these sequences can be used to generate the most probable sequence (consensus sequence) which can then be used to fish out similar sequences. The parameters in the HMM profile can then be adjusted again by incorporating more sequences of a particular class of biological significance.

Tools

1. SAM: Sequence Alignment and Modeling, developed at University of California, Santa Cruz by David Hausseler.
   Website: http://www.cse.ucsc.edu/research/compbio/sam.html

2. HMMer: developed by Sean Eddy at Washington University, St. Louis. This software package was used to develop the Pfam database of protein families using HMM profiles for these families.
   Website: http://hmmer.wustl.edu/

Ref:

1. Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, 235:1501-1531.
2. Eddy,S. R. (1996). Hidden Markov models. *Curr. Opin. Struct. Biol.*, 6:361-365.
3. Gribskov, M., Luthy, R., and Eisenberg, D. (1990). Profile analysis. *Meth. Enzymol.*, 183:146-159.

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG consists of Gene Catalog (GENES) and a graphical database of biochemical networks (PATHWAY). The pathway database consists of metabolic network maps and regulatory network maps. The metabolic maps for various organisms are generated by using a set of Basic reference maps onto which the organism specific maps is plotted automatically by matching the EC numbers in Gene Catalog. On the other hand, regulatory network maps are too divergent and therefore are generated manually. Gene cataloging and EC number assignment is done automatically and is accepted after manual verification in GENES database. KEGG is a resource geared towards functional bioinformatics and therefore, KEGG is being used in the functional annotation of genes and discovery of biochemical networks.

Ref:

Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono and Minoru Kanehisa.(1999)KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acid Research* Vol 27: 29-34

Website: http://www.genome.ad.jp/kegg/kegg2.html

MOTIF

MOTIF is a web resource developed as part of GenomeNet initiative to search Domains in a protein sequence using various specialized databases like: Pfam, BLOCK, Prosite etc. Optimum results from MOTIF are integrated into the KEGG metabolic network maps. These results may accessed by clicking on the enzyme #, motif results are available in the SSDB (Sequence Similarity Database) section.

Website: http://motif.genome.jp/

PFAM: Protein Families

Pfam is a database of protein families assembled using multiple sequence alignments and profile HMMs. Therefore, Pfam is primarily used to identification and annotation of function to various proteins. Pfam is often the first step towards homology modeling. Using Pfam the probable domains in a sequence can be identified.

Ref:

Alex Bateman, Ewan Birney, Richard Durbin, Sean R. Eddy1, Kevin L. Howe and Erik L. L. Sonnhammer (2000). The Pfam Protein Families Database. *Nucleic Acid Research* Vol 28: 263-66

Website: http://www.sanger.ac.uk/Software/Pfam/

Protein Structure Prediction Server (PSIPRED)
PSIPRED incorporates three prediction methods, a highly accurate secondary structure prediction method; MEMSAT 2 a transmembrane topology prediction method and GenTHREADER, a sequence profile based fold prediction method. The results are sent via email.
Website: http://bioinf.cs.ucl.ac.uk/psipred/psiform.html

SCOP: Structural Classification of Proteins
SCOP is an structural classification of proteins developed by Murzin and co-workers (1995). The fundamental unit of classification in SCOP is the protein domain. As a result, all the protein domains are classified hierarchically into family, superfamily, folds, classes.

40452 Domains organized into:
- 2327 Families
- 1294 superfamilies
- 800 folds

Correspond to 20619 Protein Data Bank entries.

SCOP classification has been used to develop automatic classification methods like the SuperFamily database. It has been used extensively to understand enzymatic function evolution, in the study of sequence and structure variability and its dependence in homologous proteins.
Ref: Murzin, A. G., Brenner, S. E., Hubbard, T.J.P., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540.
Website: http://scop.mrc-lmb.cam.ac.uk/scop

SUPERFAMILY
Superfamily is a set of HMM profiles built using SAM for the various superfamilies identified by SCOP. Therefore, the profiles can be used to identify distant homologous and domains in proteins.
Ref: Gough, J., Karplus, K., Hughey, R. and Chothia, C. 2001. Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. *J. Mol. Biol.*, 313(4), 903-919.
Website: http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/