Subject : Bioinformatics

Lesson : Introduction to Bioinformatics

Lesson Developer : Sandip Das

College/Department : Department of Botany, University of Delhi

Table of Contents

Introduction

Biological sciences, traditionally, was involved primarily with the observation and descriptive study of organisms. This approach, over a period of time, gave rise to several subject areas that amassed large amounts of factual information on morphology, inheritance, anatomy, taxonomy, life cycle, physiology, ecological and environmental relationships and infectivity. Over a period of time, the scientific community became curious to know the "basis" for these characteristic features of living organisms and variations that exist among them. This shift in scientific paradigm prompted an in-depth understanding of the "molecular" basis of life forms. Beginning from identification of the genetic material (nucleic acids) to sequencing the entire genomes of several organisms, biology has now been substantially re-defined. In this endeavor, biologists were benefited immensely by inputs from physical, chemical and mathematical sciences. The study of biological systems with a "*Why is it so*?" approach gave birth to several new areas of research viz., molecular genetics, genomics, proteomics, recombinant DNA technology, transgenic technology, etc. Extensive work in these areas on different biological systems led to the generation of large volumes of data on linkage maps, genomes, transcriptomes, proteomes and molecular structures, analysis of which became impossible using manual approaches. Use of computational power to analyze biological data was increasingly felt to be an unavoidable option leading to the birth of a new science called "Bioinformatics".

Why Bioinformatics

Imagine yourself trying to solve a complex mathematical calculation or trying to find a pattern in a jumbled up string of alphabets or numbers all by yourselves without the aid of any computational devices such as calculators or computers. Not only can such a task become extremely time consuming but may even turn out to be "unsolvable". However, if you are to have a calculator or a computer for your help, the given task may be performed in a much shorter duration of time. Of course, you need to know how to operate the calculator/computer and the sequence of commands to be given to the machine! In an analogous scenario, understanding the meaning of just four letter of life, namely Adenine, Guanine, Cytosine and Thymidine (Uracil) as building block of life and storehouse of information can prove daunting, unless we are able to decipher the hidden meaning for the

maintenance and functionality of the genome. A, G, C and T/U represent just one level of information content, and as we are familiar with the central dogma of Life, serves as the blueprint with message being conveyed from genetic material (DNA/RNA) to messenger RNA and eventually to proteins. Therefore at the minimum level, four nucleotides and twenty amino acids hold the entire key to life (we are not even discussing about the enormous variety of metabolites, biomolecules and other compounds that play a major role in functioning of Life).

Bioinformatics, therefore, attempts to unravel the genome information and can be understood to be comprising of two components:

Biology (**bio**) + Information Technology (**informatics)** = Computational Biology

It can be summarized as the use of information technology to generate, acquire, manage and analysis data related to biological sciences.

Computer and internet have played a major role and may be taken as the backbone on which the entire field of bioinformatics is flourishing.

Algorithms or computers programs are specialized programs/softwares written by specialists consisting of a well-defined set of steps for generation, storage and analysis of data.

The need for of development of high speed processing or computing of biological data was felt primarily on the account of the huge volume of sequencing data that was being generated. In a matter of 10 years, the cost of sequencing has dropped from nearly US$5200.00 per megabase in September 2001 to currently at 0.09cents per megabase in January 2012 (http://www.dnasequencing.org/history-of-dna).

From a few hundred megabases/year based on Sanger's di-deoxy chain termination method of sequencing, today we can generate close to 6 billion bp/ two weeks using one of the Next Generation Sequencing machines (http://www.dnasequencing.org/history-of-dna; http://www.illumina.com/systems/hiseq_comparison.ilmn), the need for even higher performing computational tools are even greater!

Although bioinformatics is largely concerned with analysis of biological data using computational tools, it may be added that it has rapidly emerged as a multidisciplinary science that touches upon subject areas in all branches of science, including physical sciences, chemical sciences, mathematics, artificial intelligence and so on.

Today, bioinformatics can be applied to analysis of a variety of data and some of these are as given below:

- **DNA sequence:**
  - Annotation
  - Analysis such as
    - Similarity search
    - functional information,
    - evolution,
    - polymorphism,
- **RNA level:**
  - Expression analysis using
    - Microarray
    - RNA sequencing
  - Structure prediction
- **Protein level:**
  - Domain and motif analysis
  - Structure determination
  - Evolution
  - Functional role
- **Whole genome/cell/tissue/organism level:**
  - Genome structure and comparative genomics
  - Interactome analysis
  - Metabolic pathways
- **Drug design**

The key to successful implementation of bioinformatics tools and their application is to organize the massive volumes of data in a user-friendly and easily accessible manner. The following section will introduce you to a few representative databases from the areas that have been listed above:

Databases and tools

Biological databases serve a critical purpose in the collation and organization of data related to biological systems. They provide computational support and a user-friendly interface to a researcher for meaningful analysis of biological data viz., gene and protein sequences, molecular structures, etc. In the recent past, computational tools and techniques have also been successfully used for simulation studies on biological macromolecules, their structures and interactions, molecular modeling and drug design accumulating significant amount of data in these interdisciplinary areas which would be dealt with separately in later units of this paper.

This section would provide a brief overview of different types/categories of databases. It would however, avoid detailed descriptions that can be accessed from several standard Bioinformatics textbooks or from the home pages of various databases. A few practice exercises for access and retrieval of information are provided at the end of the unit. Some of these exercises would be supported with step-by-step instructions for the benefit of beginners while others are to be completed by students on their own.

Questions:

How would I know whether a database relevant to my interest/study exists or not?

How can I be assured of the authenticity of the information available in any database?

Answer:

The journal, Nucleic Acids Research (NAR), publishes in its January issue every year, a comprehensive compilation of all *peer-reviewed* databases and online tools.

These issues can be accessed at http://nar.oxfordjournals.org/. The **peer review** process ensures that the published literature and its contents are accurate.

Bioinformatics based Database and tools for DNA analysis:

a. Annotation is one of the first steps and generally refers to adding identifying features to the sequence. Generally, annotation can be performed by simply comparing an unknown sequence with a DNA sequence, which has already been annotated. This is carried out based on the principles of sequence comparisons with a hypothesis that if two sequences share similarity, then they should also share characteristic features. However, genome annotation or identifying features of genomes are more challenging. For example, genome annotation can be subdivided into two categories

- **Comparative genomics** based (i.e. based on comparison with other genomes)
- ***Ab-initio*** based (i.e. from the beginning)

Depending on the source of the sequence, annotation tools are designed for either prokaryotic genomes or eukaryotic genomes.

GeneMark.hmm (Lomsadze et al. 2005; http://exon.gatech.edu/index.html; **GeneScan** (http://genes.mit.edu/GENSCAN.html), and **FGENESH** (http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind) are comparative genomics based eukaryotic gene prediction tool. They take unknown DNA sequence as input sequence, compares with several stored genome features in its database and can predict

- presence of genes
- position of genes
- strand on which genes are present,
- exon positions
- translation product

Some other softwares developed for gene prediction include **Artemis** by the **EBI-EMBL** (http://www.sanger.ac.uk/resources/software/artemis/).

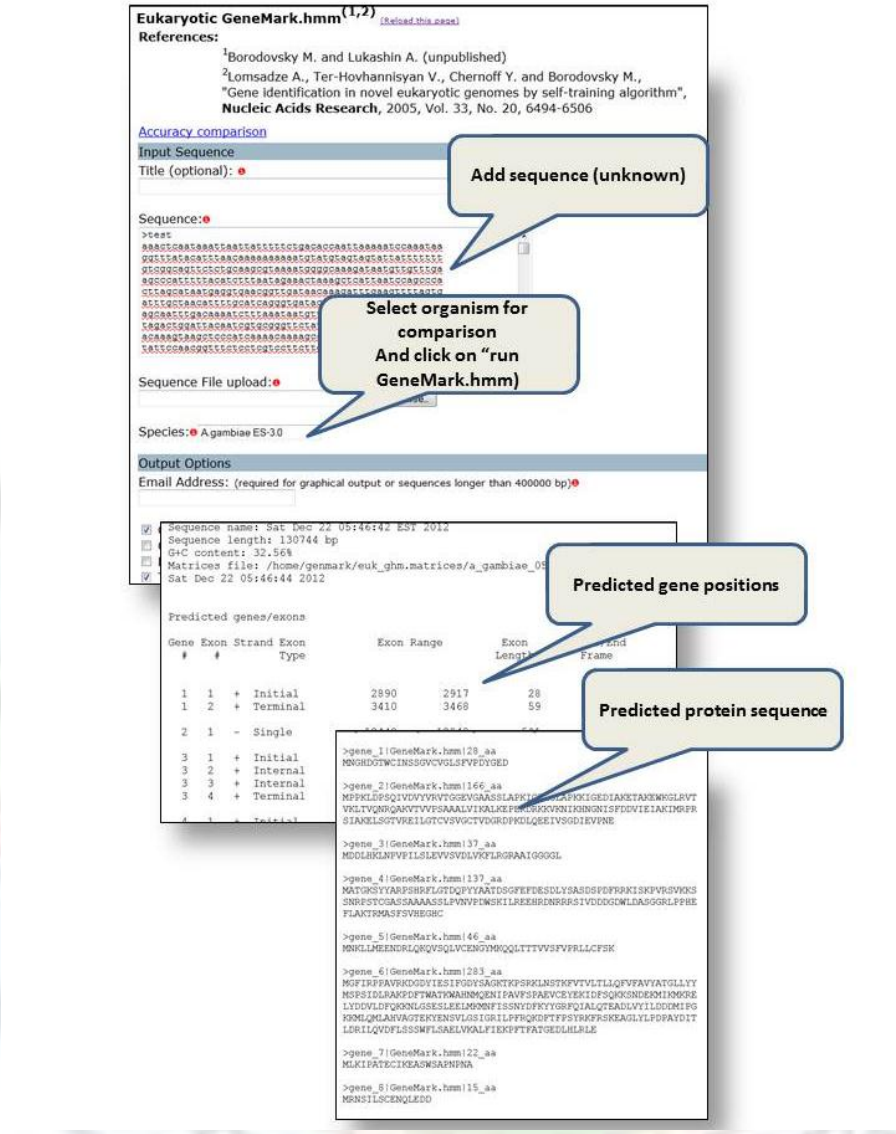


Figure: Genemark.hmm, eukaryotic gene prediction tool based on comparative genomics.

Source: http://exon.gatech.edu/eukhmm.cgi

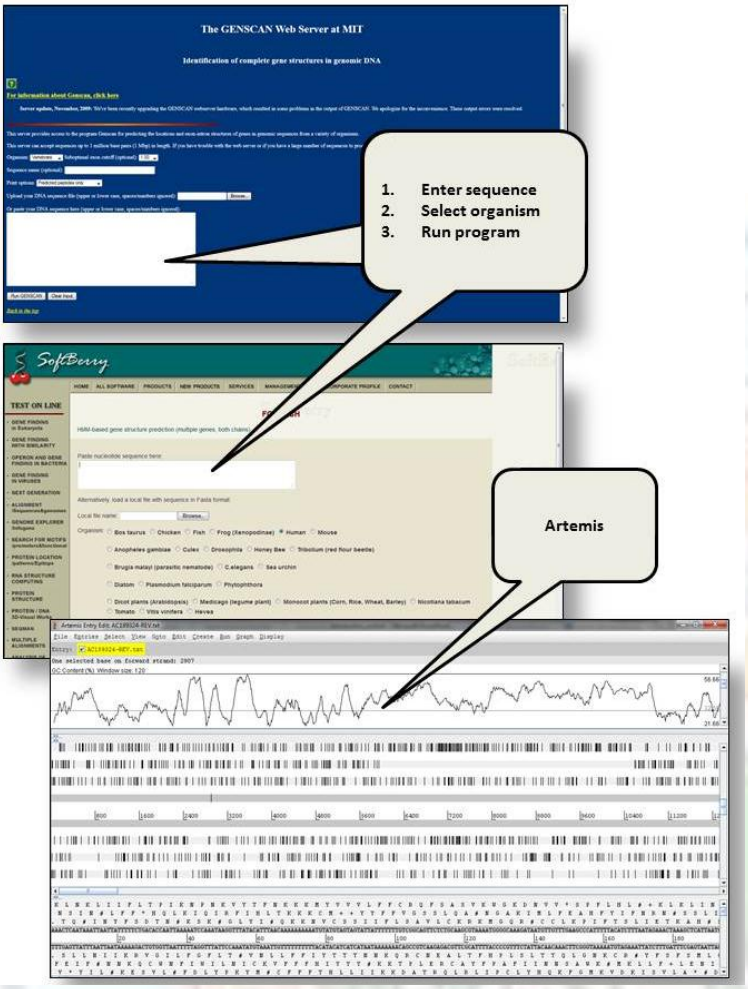


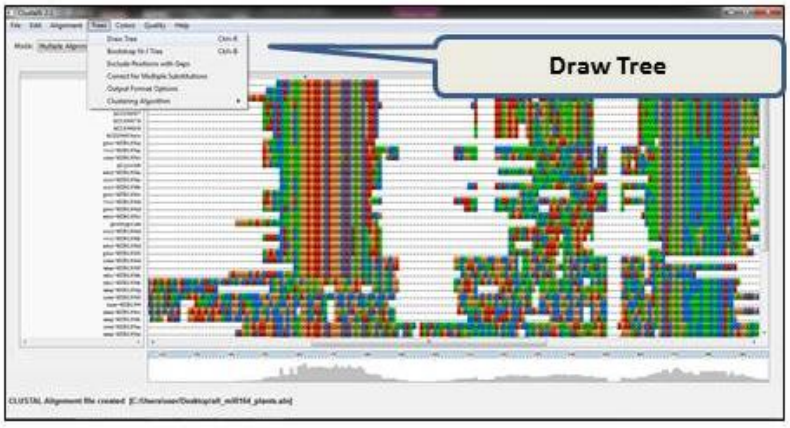
Figure: Gene prediction tools- Genescan, FGENESH and Artemis (top to bottom)

Source: http://genes.mit.edu/GENSCAN.html
http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind
http://www.sanger.ac.uk/resources/software/artemis/

b. Similarity searching can be performed using the BLASTN tool at NCBI (http://www.ncbi.nlm.nih.gov) and a detailed description in available in the chapter on NCBI BLAST.

c. Molecular evolution using DNA sequence can be performed using multiple sequence alignment. The most common tool for multiple sequence alignment is Clustal that can either be used as a web-based service or the software can be downloaded from http://www.clustal.org/. It employs progressive alignment as to perform a MSA, Clustal first

creates a global pairwise alignment for all sequence pairs with alignment/similarity scores and then starts the MSA with the two sequences with highest score and progressively adds more and more sequences to complete the alignment. The MSA can be further analysed using software such as MEGA to reveal evolutionary relationship. For details on how to construct multiple sequence alignment and phylogenetic tree please refer to the chapter on molecular phylogeny and Multiple Sequence Alignment



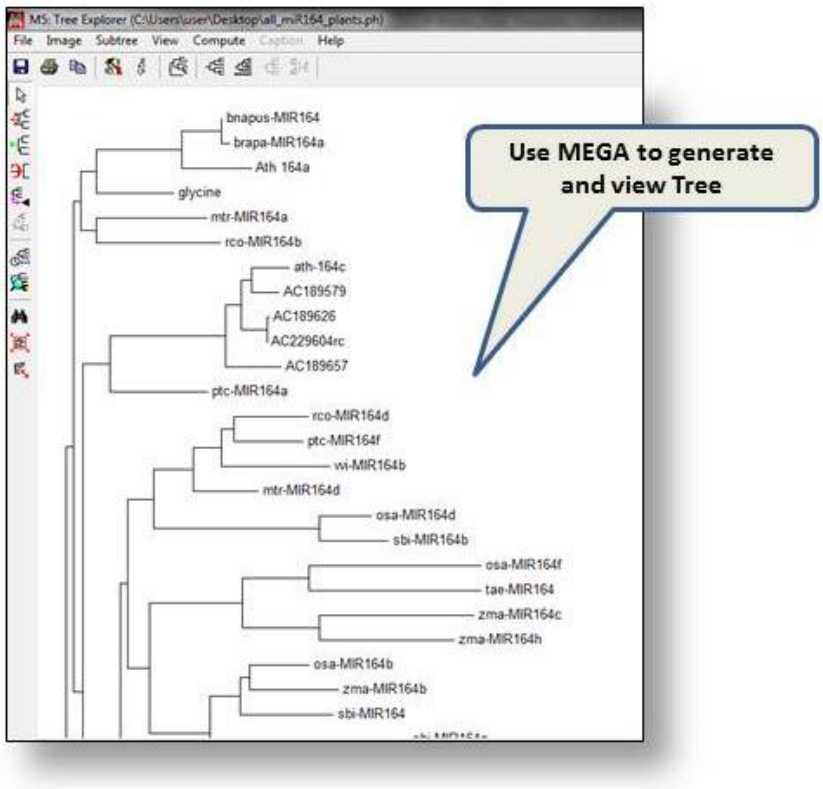


Figure: Computational tools such as ClustalX/ClustalW or Clustal Omega together with MEGA (http://www.megasoftware.net/mega.php) can be used for molecular phylogeny
Source: http://www.megasoftware.net/mega.php

Bioinformatics based databases and tools for RNA analysis:

a. Gene Expression analysis:

Gene Expression Omnibus (GEO) is a central repository at NCBI that catalogues expression profile information from a variety of organism(http://www.ncbi.nlm.nih.gov/geo/). A similar database at EMBL is maintained as **Arrayexpress** (http://www.ebi.ac.uk/arrayexpress/). Specialized databases that are specific to organisms include **eFP browser (**http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi) and **AtGenExpress** (http://www.weigelworld.org/resources/microarray/AtGenExpress/) for *Arabidopsis thaliana*, **Modmine** for *Drosphila melanogaster* (http://intermine.modencode.org/query/portal.do?class=Project&externalids=Celniker) is maintained under the Berkley Drosophila genome project (http://www.fruitfly.org). Brain atlas is a database for expression datasets in brain tissue, irrespective of the organism.

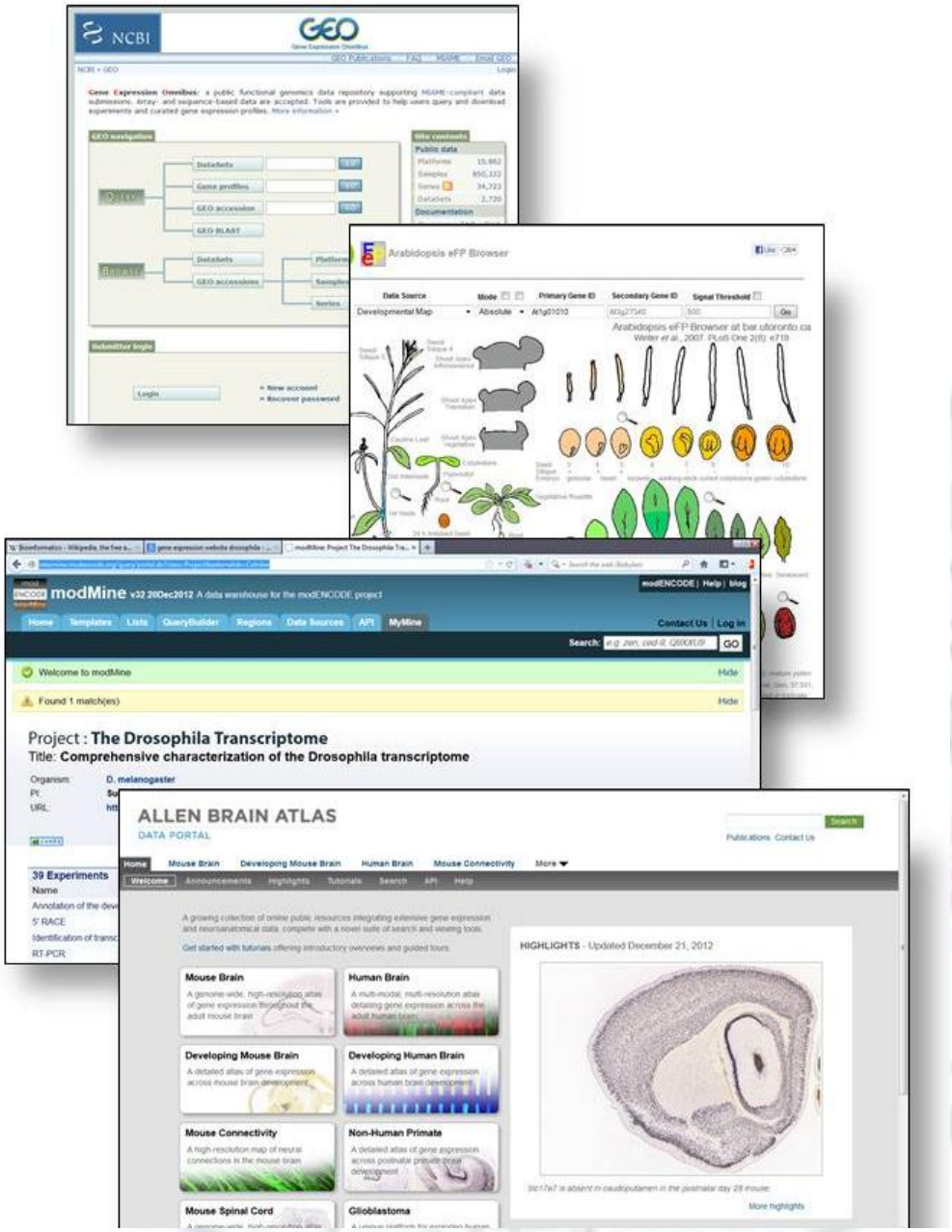


Figure: Representative webshots of Gene expression databases (GEO, eFP browser, Modmine and Brain-Atlas)

Source: ,http://www.ncbi.nlm.nih.gov/geo/ , http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi , http://www.fruitfly.org/ , http://help.brain-map.org/display/humanbrain/Allen+Human+Brain+Atlas

b. **RNA structure prediction:** A large number of RNA molecules undergo 3-D structural conformation to obtain functionality. Bioinformatic tools and databases have been created that are useful for prediction of RNA structures.

One software that is commonly employed for RNA structure prediction is mFold (http://mfold.rit.albany.edu/?q=mfold), which accepts nucleotide sequences and predicts the all the probable structures that can be derived from the sequence. Some other RNA fold bioinformatics tools include RNAFold webserver (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi).

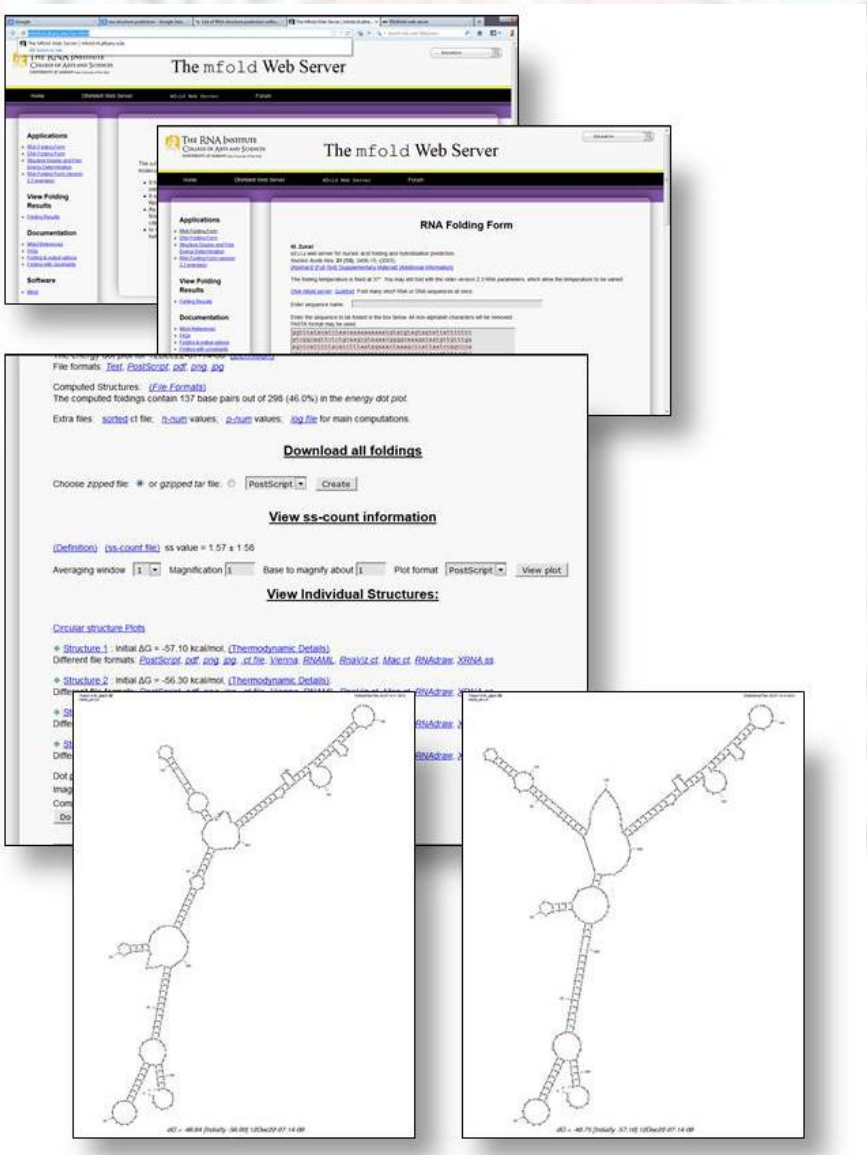


Figure: RNA folding tool, mFold

Source: (http://mfold.rit.albany.edu/?q=mfold)

Bioinformatics based database and tools for protein analysis:

a. Sequence analysis: Similarity searching using protein sequences are performed using **BLASTP** analysis tool at the NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&SHOW_DEFAULTS=on&LINK_LOC=blasthome) and a detailed description of BLASTP is provided in the section on BLAST.

Several online bioinformatics tool for protein analysis have been made available at Expasy (http://www.expasy.org/). Some of these tools are:

i. **PANDIT Plus:** for resources related to domain structure and protein families (http://panditplus.org/). Another such tool and database is the **PROSITE** (http://prosite.expasy.org/)

ii. **ProtParam:** This tool has been developed for prediction of physical (such as Molecular weight) and chemical (such as pI) properties of proteins (http://web.expasy.org/protparam/)

iii. **PSORT:** A tool for predicting the sub-cellular localization of proteins (http://www.psort.org/)

iv. **Coiled–Coil Prediction:** for prediction of coiled-coil regions (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_lupas.html)

b. Structure prediction of proteins: Protein structure prediction is one of the major challenge areas in bioinformatics. Most of the structures that have been solved accurately are based either on **NMR spectroscopy** or **X-ray crystallography.** Although there is information available that the primary sequence gives rise to secondary which in turn folds into tertiary and eventually quaternary structure, no precise algorithm has been developed that can predict all protein structures. Using computational tools, three strategies have been proposed for protein structure prediction,

i. **Homology modeling**

ii. **Fold recognition** or **Threading,** and

iii. Ab-initio prediction

SWISS-MODEL Workspace http://www.expasy.org/ is a homology based protein structure prediction program.

PHYRE2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index) is a fold recognition based protein structure prediction tool

HMMSTR/Rosetta (http://robetta.bakerlab.org) and now available as ROBETTA (http://robetta.bakerlab.org/) is an ab-initio protein structure prediction tool

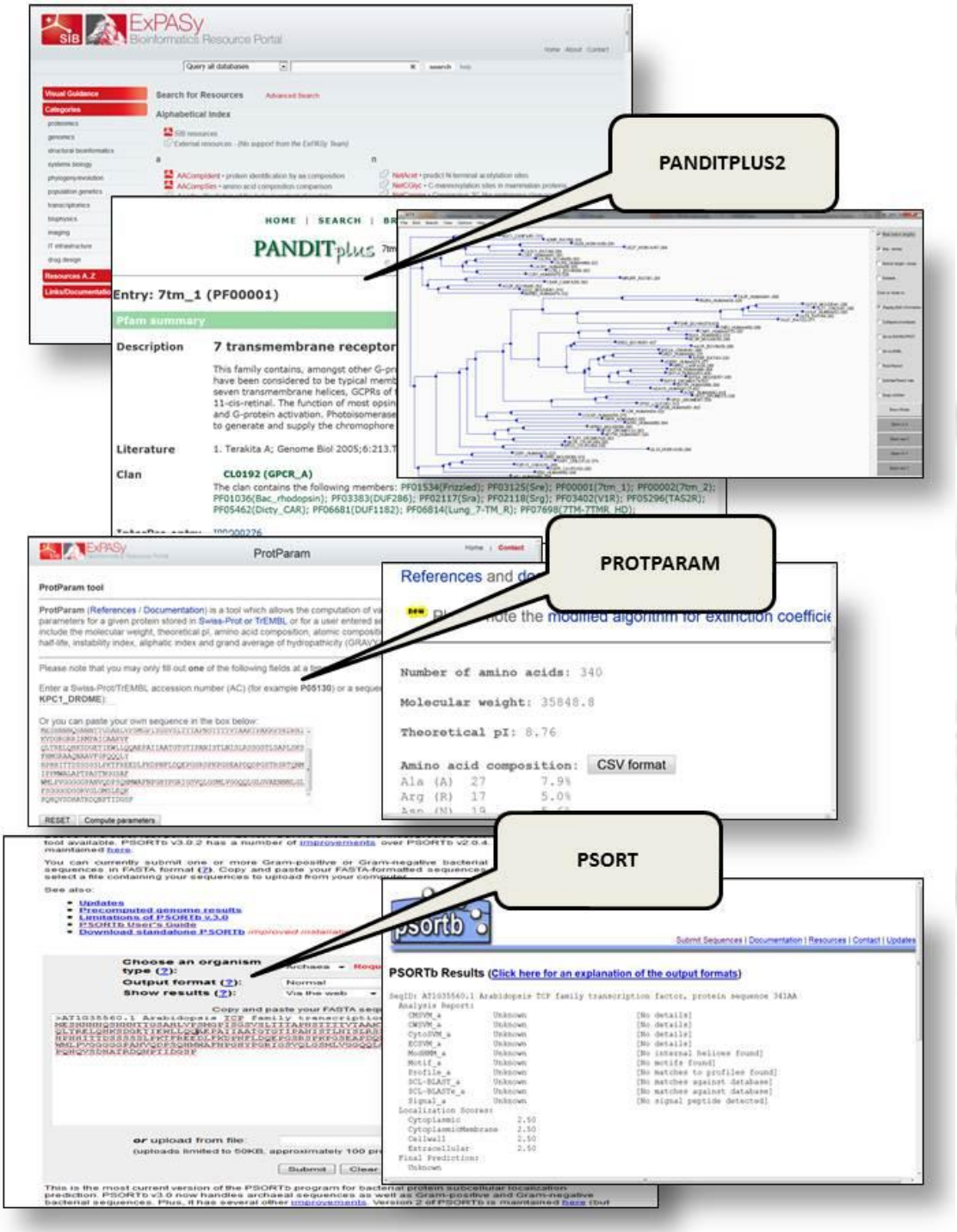


Figure: Expasy server hosts several bioinformatics tools and databases for analysis of protein sequence

Source: http://www.expasy.org/ , http://panditplus.org/, http://web.expasy.org/protparam/ , http://www.psort.org/

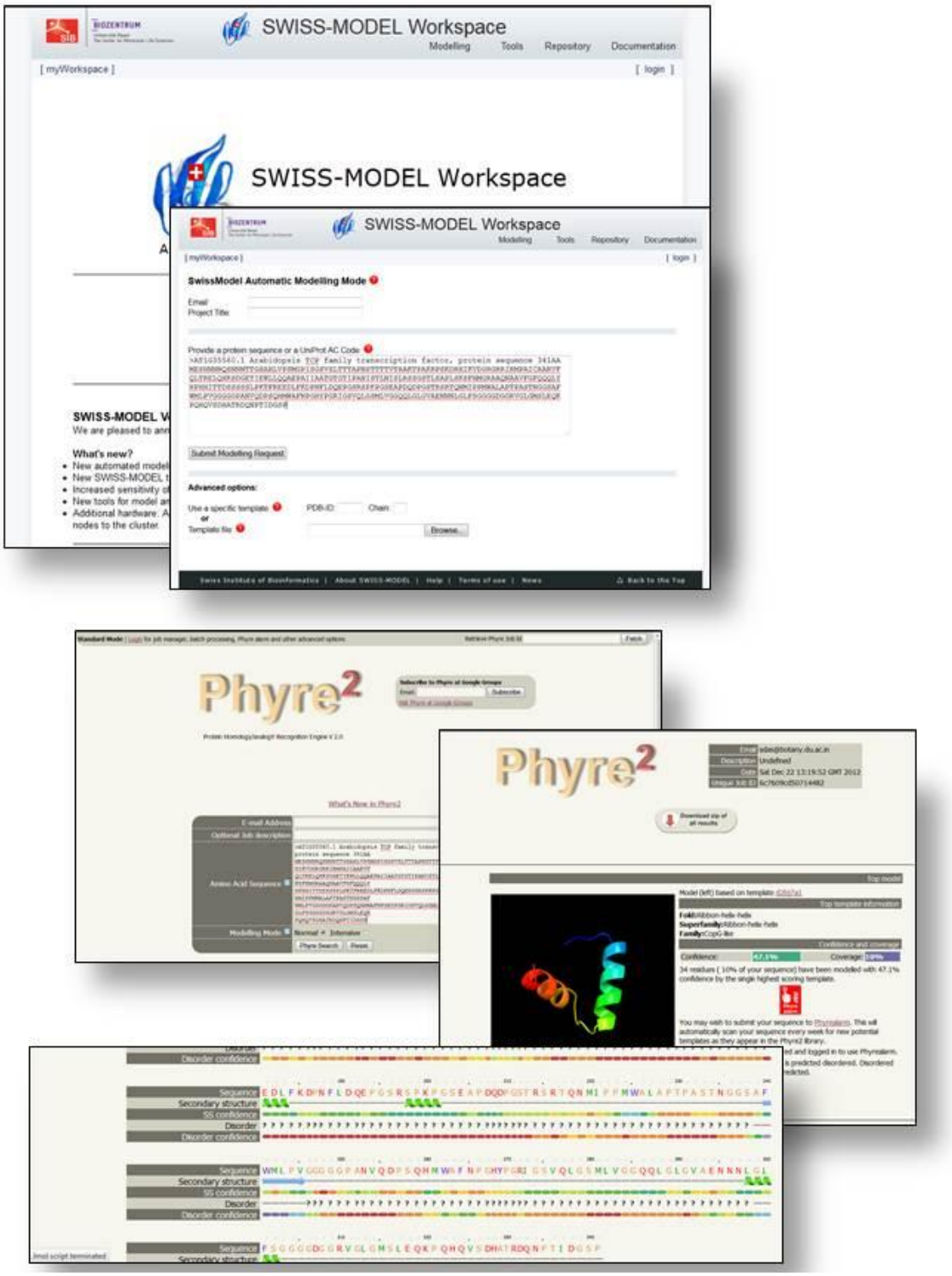


Figure : Protein structure prediction webserver based on homology modeling (Swiss-Model) and Threading (PHYRE2)

Source: http://swissmodel.expasy.org/workspace/

http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index

Bioinformatics based databases and tools for whole genome analysis:

a. Comparative genomics and genome structure:

The last few years has seen the completion of high quality draft genome sequences from a wide range of organism. The Genome database at NCBI currently lists total of 3468 viral genomes, 15360 from prokaryotes and 2268 from eukaryotes (as on December 22, 2012; http://www.ncbi.nlm.nih.gov/genome/browse/). The availability of genome sequences from such a variety of organism has allowed researchers to reconstruct the genome structure at the highest resolution possible, i.e. at the level of nucleotides; and also allowed comparison of entire genome rather than just a few kilobases. In other words, sequencing has enabled researchers to not only get a bird's-eye-view of the genome but also permits them to zoom into virtually any region of the genome. Comparison of the genomes to itself allows us to understand the overall genome organization and composition such duplication and inversions; whereas comparison of two or more genomes permit us to understand the evolution of genomes vis-à-vis each other.

For example, the **Pairwise Sequence Comparison (PASC; http://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi?textpage=overview; Bao et al. 2008)** at NCBI is a tool for comparison of sequences within viral genomes.

TaxPlot (http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi;) allows users to compare genomes in a three way plot that identifies protein homologs present across genomes. The tool is available for genome-wide comparisons across microbial and eukaryotic genomes.

Global alignment tools such as AVID (http://pipeline.lbl.gov/cgi-bin/gateway2) that allow users to compare large segments or entire genomes from various organisms have become a powerful tool for comparisons across genome irrespective of their phylogenetic distance or coding potential (as compared to PASC which is limited to viral genomes and TaxPlot that compares protein homologs).

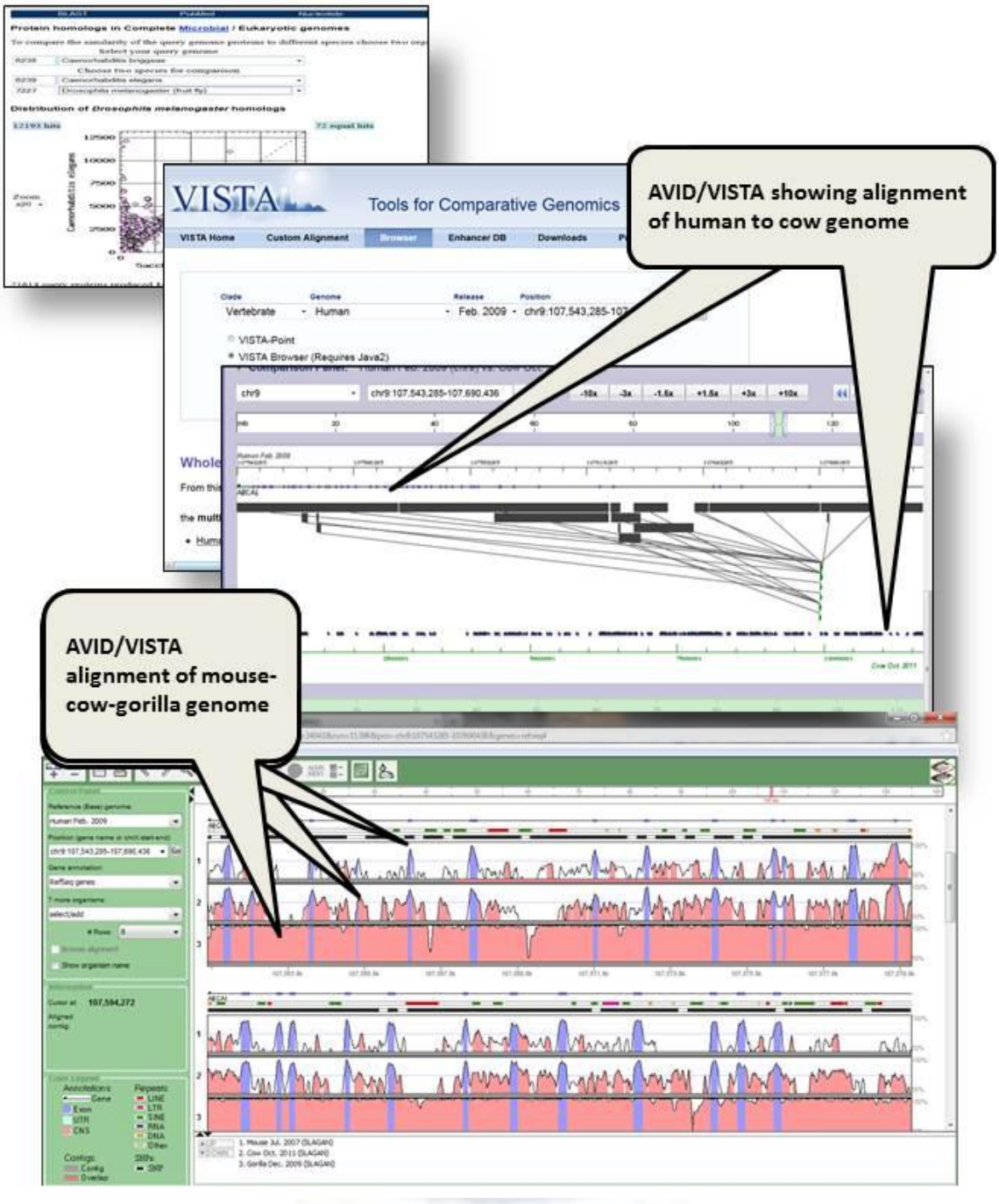


Figure: Tools of comparative genomics such as TaxPlot and AVID allow users to compare genomes to understand homology, organization, and evolution across genomes

Source: http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi , http://pipeline.lbl.gov/cgi-bin/gateway2

Dotplots, that represent comparisons of genomes or sequences in a X- and Y-axis plot can be used to reveal genomic changes such as duplication, inversion, deletions. Such analysis can be performed within the same genome or can also be across two genomes. **PipMaker** and **Multipipmaker** (http://pipmaker.bx.psu.edu/pipmaker/), **CoGe-SynMap** (http://genomevolution.org/wiki/index.php/Syntenic_dotplot), **GBrowse–Syn** (http://gmod.org/wiki/GBrowse_syn) all allow comparisons to be performed between various genomes.

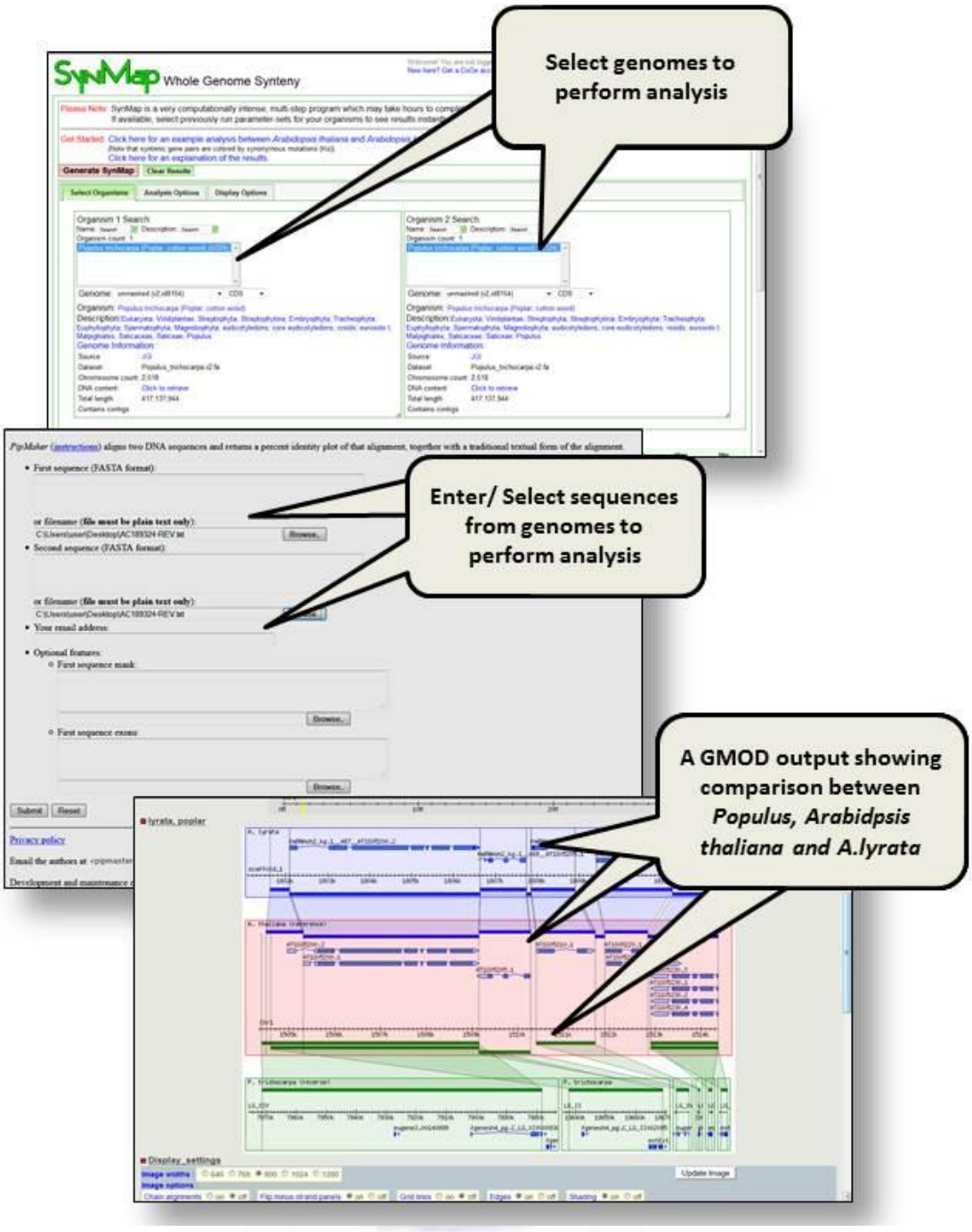


Figure: Comparative genomic tools

Source: http://genomevolution.org/CoGe/SynMap.pl, http://gmod.org/wiki/GBrowse_syn

b. Interactome and Biological Network tool: Interactome can be defined as a discipline of proteomics / genomics that aims to reveal the direct and indirect interactions within and between proteins and other cellular macromolecules in a cellular environment. Drawing up such an **interactome** map helps us understand the regulation and functioning of the genome. **IntAct** (http://www.ebi.ac.uk/intact/) available at EBI-EMBL, together with CytoScape (visualization tool) is one such tool that has nearly 304500 interaction datasets. **Database of Interacting Proteins** (http://dip.doe-mbi.ucla.edu/dip/Main.cgi),

Apart from protein interaction data, KEGG (Kyoto Encylopedia of Genes and Genome; http://www.genome.jp/kegg/pathway.html) is a database cum analysis tool for a wide range of metabolic pathways for biosynthesis and degradation of biological compounds.

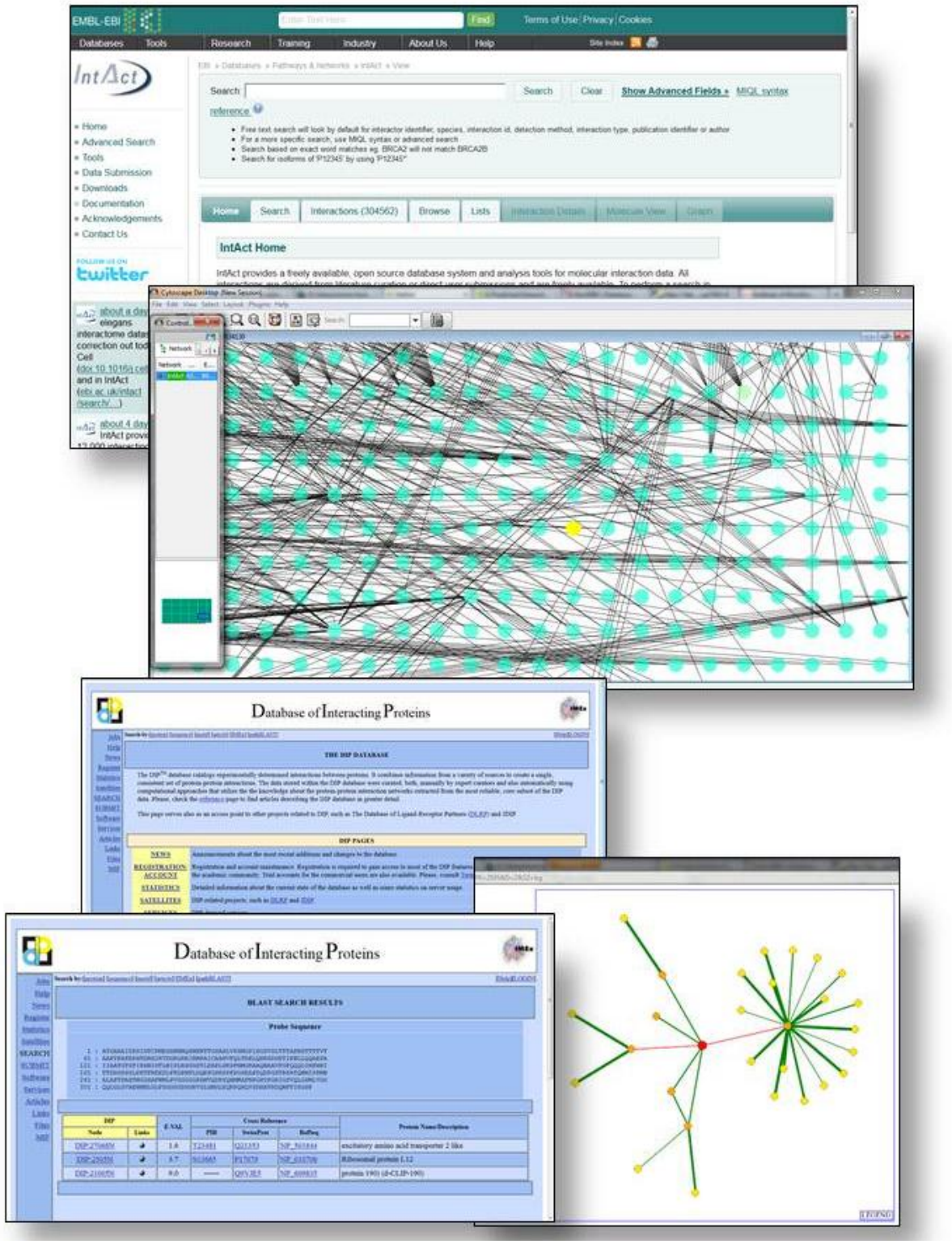

EMBL-EBI
Databases
Tools
Research
Training
Industry
About Us
Help
IntAct
Search
Clear
Show Advanced Fields »
MIQL syntax
Home
Advanced Search
Tools
Data Submission
Downloads
Documentation
Acknowledgements
Contact Us
Interactions (304562)
Browse
Lists
IntAct Home
Database of Interacting Proteins
THE DIP DATABASE
DIP PAGES
NEWS
REGISTRATION/ ACCOUNT
STATISTICS
SATELLITES
Database of Interacting Proteins
BLAST SEARCH RESULTS
Probe Sequence

Figure : Interactome database such as IntAct and DIP reveal network of interactions
Source: http://www.ebi.ac.uk/intact/, http://dip.doe-mbi.ucla.edu/dip/Main.cgi

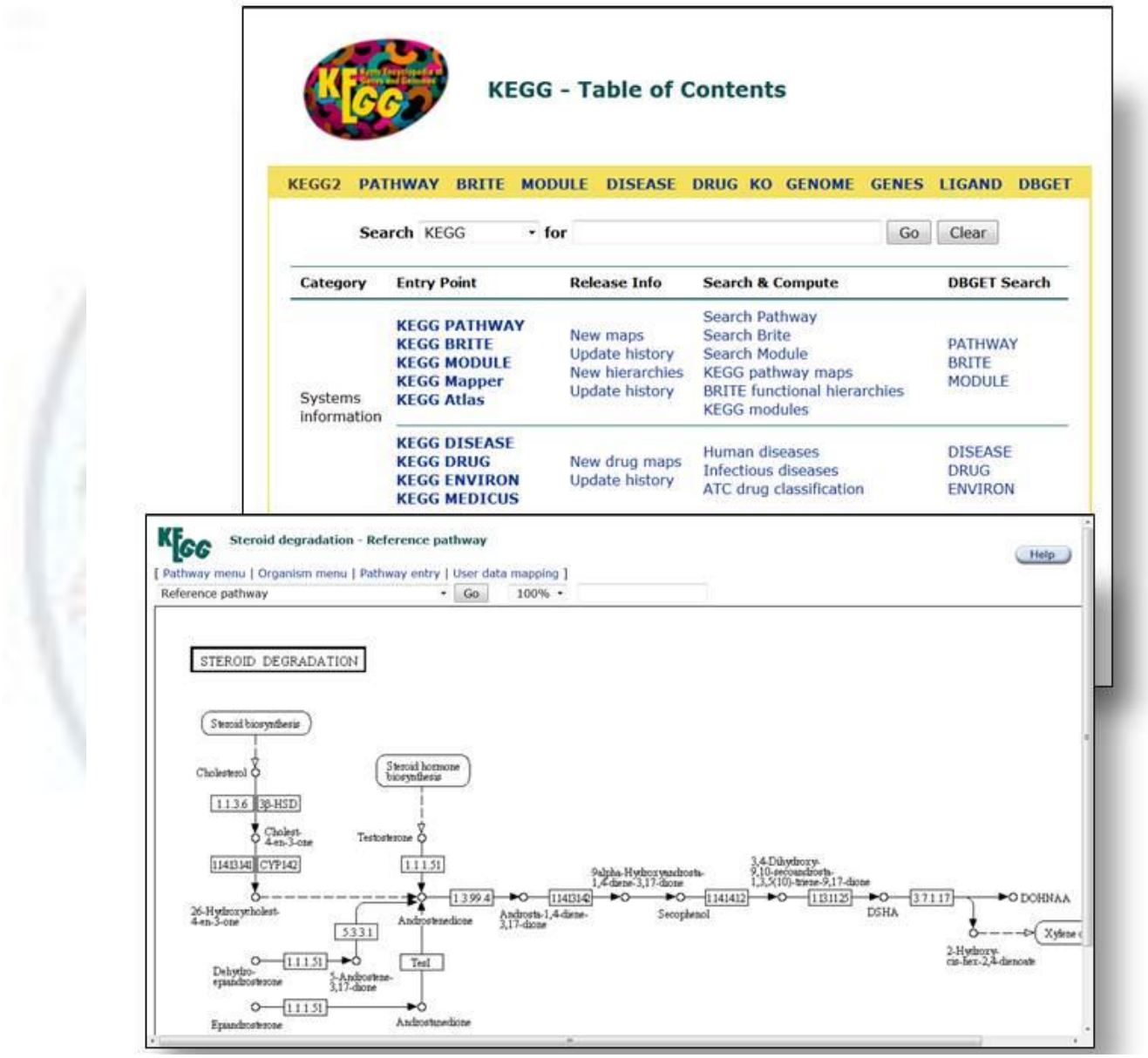


Figure : KEGG database store information about metabolic pathways
Source: http://www.genome.jp/kegg/pathway.html

Application

The preliminary role of bioinformaticians was to organize the available data into a format that is easy to access, visualize and analyze. Therefore, the initial thrust area in bioinformatics was the development of Databases, which served as electronic filing cabinets for biological data. Subsequently, computational power was finding answers to the Living organisms have been subjected to innumerable studies at various levels viz., structure (morphology, anatomy), function (physiology, biochemistry), inheritance (genetics), evolution, taxonomy, etc. to name a few. Scientific research over the last several decades, in addition to the above approaches, has also attempted to unravel the molecular basis of processes that are integral to organism biology. These studies were initially focused on a subset of relatively less complex organisms that came to be popularly referred to as Model Organisms or Model Systems. Such organisms belonged to a wide range of life forms ranging from viruses and bacteria to higher plants and animals. Notable examples include *Drosophila*, *C. elegans*, Arabidopsis, mice, yeast and more recently *Oryza sativa*, *Medicago*, *Lotus*, etc. Molecular genetic studies on many of these life forms led to the development of markers and linkage data, which in turn, facilitated whole genome-sequencing programs to extract the encoded information (genome sequence) that supports life. Subsequent analysis of gene function based on expression profiling (transcriptome studies) and mutant analysis (functional genomics) contributed further to our understanding of biological systems. Rapid developments in sequencing chemistry ushered in an era of high throughput genome and transcriptome sequencing, which led to a virtual explosion of biological data across the world transgressing the limits of "model systems" for biological studies. Seminal developments in Bioinformatics centered mainly on the development of Databases, which functioned as electronic filing cabinets for the organization and analysis of large amounts of biological data that were generated from a variety of such studies.

Some of the major application areas of bioinformatics are

a. Molecular medicine using Drug design, gene therapy,
b. Waste cleanup using information gained from microbial genomes
c. Generation of clean energy
d. Design of Bio-catalysts and improved strains of bacterial strains for production of industrially useful products such as antibiotics, enzymes, metabolites

e. Improved and better, designer crops such as with increased adaptability to altered climatic conditions, high yielding, production of edible vaccines, improved nutritional characteristics

Summary

This lesson introduces you to the world of bioinformatics albeit in an abridged manner. The deluge of biological data generated as a result of high-throughput technologies can be maintained and analysed only via a computational approach or bioinformatics. The advent of internet and high speed computational power has been a powerful ally in the rapid development and acceptance of bioinformatics tools. Bioinformatics can be applied to analysis of various categories of data ranging from DNA sequence, evolution, gene expression, RNA structure prediction, and protein sequence-structure analysis. Apart from these, bioinformatics is necessary for the functional characterization at whole genome and organism level and fields such as Comparative genomics, Interactome and Metabolome analysis have been developed. Finally, bioinformatics is playing a pivotal role in various application areas such as drug design, waste cleanup, disease diagnosis, generation of clean energy, bio-fuel and bio-catalysts, genetically modified microbes and plants for better yield, adaptability and increased/altered nutritional values.

Exercises

1. What are the major branches of bioinformatics?
2. Comment on the various databases and analysis tools of DNA.
3. Bioinformatics comprises of ------ + ----------.
4. Which journal publishes a yearly issue on bioinformatics databases?
5. What are the various approaches to annotation?
6. Why should you perform annotation of sequence?
7. Prepare a list of softwares and tools for genome annotation. With the help of a flowchart, list the steps of annotation using Genemark.hmm and GeneScan.
8. Comment on the various databases and analysis tools of RNA
9. Can sequence information be employed to understand evolution? Mame some tools for such an analysis?
10. What tools would you choose to perform gene expression analysis? What are the features associated with any two such tools?

11. RNA structure can be predicted using ----------.
12. Comment on the various databases and analysis tools for proteins.
13. Write a concise account of protein similarity search tool.
14. How are protein structures determined?
15. What are the approaches for protein structure prediction?
16. PHYRE2 is employed for ------- based ----------.
17. Rosetta is an ---------- tool.
18. Why should you perform comparative genomic analysis?
19. What is the utility of PASC and TaxPlot?
20. What is a dotplot? What tools are available to you to perform dotplot?
21. Define Interactome. What are the different resources for interactome analysis?
22. Retrieve any metabolic pathway using KEGG database.
23. What are the various application areas of bioinformatics?
24. Define the following:
    a. Algorithm
    b. Domain
    c. Motif
25. Expand the following:
    a. BLAST
    b. EST
    c. NCBI
    d. GEO
    e. NMR
    f. KEGG

Glossary

Ab-initio: "from the beginning", a term used to denote processes or methods such as gene prediction or protein structure prediction, in the absence of any prior information or guides as in comparative methods

Algorithm: A series of steps or instruction in a computer program that are used to execute a software

Annotation: The process of adding identification features to genome, sequence, transcripts or proteins

Bioinformatics: The application of computational tools for generation, acquisition, organization and analysis of biological data

Domain: A conserved block of sequence that has a defined sequence composition and function.

Dot-plot: A method of sequence comparison that compares two sequences and displays the matches as series of dots. The two sequences form the X- and Y- axis and the similarity or differences appear as diagonal series of dots and dashes

Motif: A short sequence stretch that has a defined function. For example Asn-X-Ser/Thr is a motif that is used for attachment of carbohydrates to proteins; "WRKY" motif is found in all WRKY domain containing proteins. "TATA" box is a motif found in Type II promoters. A domain can consist of several motifs.

References

Works Cited

1. Lomsadze A., Ter-Hovhannisyan V., Chernoff Y. and Borodovsky M., "Gene identification in novel eukaryotic genomes by self-training algorithm", Nucleic Acids Research, 2005, Vol. 33, No. 20, 6494-6506
2. Dimitrieva S. and M. Anisimova. 2010. PANDIT*plus*: Towards better integration of evolutionary view on molecular sequences with complementary bioinformatics resources. Trends in Evol Biol.
3. Bao Y., Kapustin Y. & Tatusova T. (2008). Virus Classification by Pairwise Sequence Comparison (PASC). Encyclopedia of Virology, 5 vols. (B.W.J. Mahy and M.H.V. Van Regenmortel, Editors). Oxford: Elsevier. Vol. 5, 342-348.
4. Dimitrieva S. and M. Anisimova. 2009. PANDIT*plus*: Towards better integration of evolutionary view on molecular sequences with supplementary bioinformatics resources. Trends in Evol Biol.
5. Protein structure prediction on the web: a case study using the Phyre server Kelley LA and Sternberg MJE. *Nature Protocols* 4, 363 - 371 (2009)
6. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. Nucleic Acids Res. 2004 Jul 1;32 (Web Server issue):W273-9

Suggested Readings

1. Bioinformatics and Functional Genomics: 2nd Edition, Jonathon Pevsner (2009), Wiley Blackwell

2. David W. Mount. (2004) Bioinformatics**:** Sequence and Genome Analysis, 2nd ed. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

3. Campbell A.M., Heyer L.J. (2006) Discovering Genomics, Proteomics and Bioinformatics. 2nd ed. Benjamin Cummings.

Web Links

http://www.dnasequencing.org/history-of-dna

http://www.illumina.com/systems/hiseq_comparison.ilmn

http://genes.mit.edu/GENSCAN.html

http://linux1.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind

(http://www.sanger.ac.uk/resources/software/artemis/)

http://mfold.rit.albany.edu/?q=mfold

http://swissmodel.expasy.org/workspace/

http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)

http://robetta.bakerlab.org

http://www.ncbi.nlm.nih.gov/sutils/pasc/viridty.cgi?textpage=overview

http://pipeline.lbl.gov/cgi-bin/gateway2

http://pipmaker.bx.psu.edu/pipmaker/

http://genomevolution.org/wiki/index.php/Syntenic_dotplot

http://gmod.org/wiki/GBrowse_syn

http://dip.doe-mbi.ucla.edu/dip/Main.cgi

http://www.ebi.ac.uk/intact/)

http://www.genome.jp/kegg/pathway.html