Chapter 9
Applied Bioinformatics Tools

Jingchu Luo

9.1 Introduction

A hands-on course mainly for the applications of bioinformatics to biological problems was organized at Peking University. The course materials are from http://abc.cbi.pku.edu.cn. They are divided into individual pages (separated by lines in the text):

- Welcome page
- About this course
- Lectures
- Exercises
- Projects
- Online literature resources
- Bioinformatics databases
- Bioinformatics tools

This chapter lists some of the course materials used in the summer school. The course pages are being updated and new materials will be added (Fig. 9.1).

9.1.1 Welcome

Welcome to ABC – the website of Applied Bioinformatics Course. We'll learn, step by step, the ABCs of:

- How to access various bioinformatics resources on the Internet
- How to query and search biological databases

J. Luo (✉)
Center for Bioinformatics, School of Life Sciences, Peking University, Beijing 100871, China
e-mail: luojc@mail.cbi.pku.edu.cn

R. Jiang et al. (eds.), *Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China*, DOI 10.1007/978-3-642-38951-1_9,

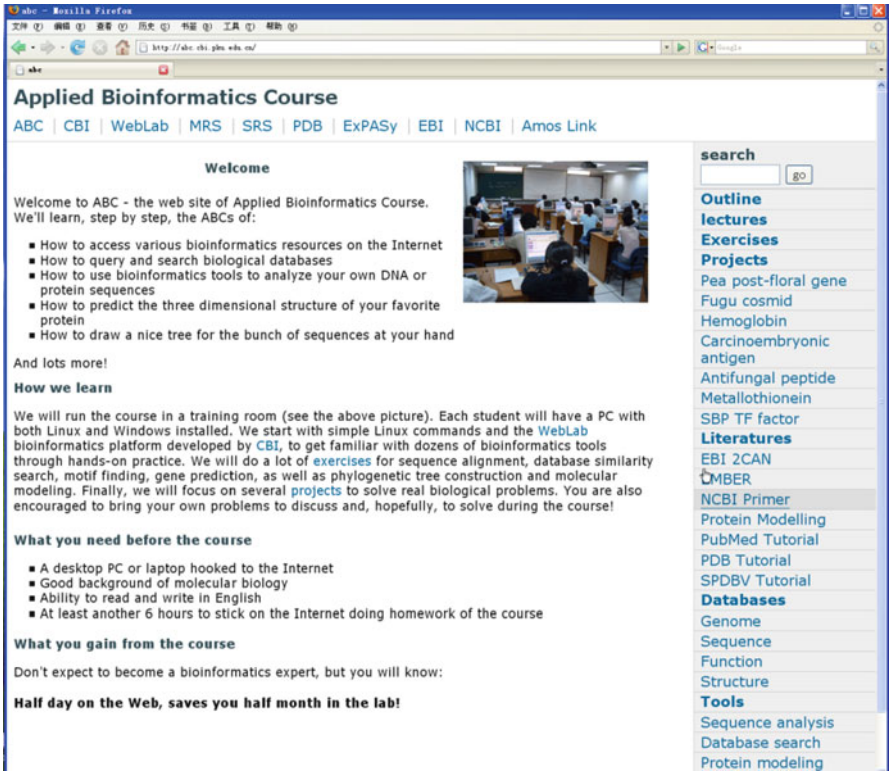

Fig. 9.1 The website of the applied bioinformatics course

- How to use bioinformatics tools to analyze your own DNA or protein sequences
- How to predict the three-dimensional structure of your favorite protein
- How to draw a nice tree for the bunch of sequences at your hand
- And lots more!

9.1.1.1 How We Learn

We will run the course in a training room (see the above picture). Each student will have a PC with both Linux and Windows installed. We start with simple Linux commands and the WebLab bioinformatics platform developed by CBI, to get familiar with dozens of bioinformatics tools through hands-on practice. We will do a lot of exercises for sequence alignment, database similarity search, motif finding, gene prediction, as well as phylogenetic tree construction and molecular modeling. Finally, we will focus on several projects to solve real biological problems. You are also encouraged to bring your own problems to discuss and, hopefully, to solve during the course!

9.1.1.2 What You Need Before the Course

- A desktop PC or laptop hooked to the Internet
- Good background of molecular biology
- Ability to read and write in English
- At least another 6 h to stick on the Internet doing homework of the course

9.1.1.3 What You Gain from the Course

Don't expect to become a bioinformatics expert, but you will know:

Half day on the web, saves you half month in the lab!

9.1.2 About This Web Site

This website is an online portal for the teaching of the semester course of computer application to molecular biology since 2000. We now use Applied Bioinformatics Course, or ABC as a simple name for this course. Here, ABC means that it is an entry level introductory course, rather than an advanced one.

9.1.2.1 Students

The aim of this course is mainly for graduate students of biology to learn how to use bioinformatics tools to solve his or her own problems. The students are mainly from:

- College of Life Sciences, Peking University (PKU)
- Graduate school, Chinese Academy of Agricultural Sciences (CAAS)

We also run training courses for graduate students and junior researchers. For example, with the help of colleagues from EMBnet (Dr. Valverde, Spanish node, and Dr. Moulton, University of Manchester, UK), two 3-day courses (30 h each) were organized for:

- Institute of Botany, Chinese Academy of Sciences (CAS)
  And two training courses (15 h each) were given for participants of:
- 2007 Summer School of Bioinformatics, Lab of Bioinformatics, Tsinghua University (TSU)

9.1.2.2 For Course Students

The course will be given mainly in Chinese, mixed with English terms. Some of the lecture slides are in English. You should have a good background of English

especially in scientific reading if you plan to study this course. Registration is needed for both PKU and CAAS students taking this as a semester course. Please see the Notice page for more details.

9.1.2.3 For Web Visitors

The web pages of this site are in English. However, the teaching materials, e.g., the lectures given by students, are in Chinese. They are in PDF format and freely available for download. For non-native Chinese speakers, however, you should have a good command of Chinese.

9.1.3 Outline

A. Getting started with applied bioinformatics

1. An overview of bioinformatics
2. The basis of molecular biology

B. Bioinformatics resources over the Internet

1. International bioinformatics centers (NCBI, EBI, ExPASy, RSCB)
2. Literature references (PubMed, PubMed Central, online books, bioinformatics courses, tutorials)
3. Databases of molecular biology (genome, sequence, function structure databases)
4. Bioinformatics tools (programs, packages, online web servers)

C. Database query

1. Database query with the NCBI Entrez system
2. Database query with the EBI SRS platform
3. Database query with the MRS platform

D. DNA and protein sequence analysis

1. Sequence alignment and dotplot
2. Protein sequence analysis
3. DNA sequence analysis
4. Database similarity search (BLAST, FASTA)
5. Phylogenic analysis and tree construction

E. Molecular modeling

1. Molecular graphics and visualization
2. Molecular modeling and structure prediction

F. Projects

1. Sequence analysis of Pisum sativum post-floral specific gene
2. MDR – gene prediction and dotplot analysis of fugu multidrug resistance gene
3. Sequence, structure, and function analysis of the bar-headed goose hemoglobin
4. CEA – protein engineering of carcinoembryonic antigen
5. Structure comparison of spider toxins and prediction of antifungal peptide
6. Sequence and structure comparison of human metallothioneins
7. Systematic analysis of the Arabidopsis transcription factor family of Squamosa-promoter binding protein

9.1.4 Lectures

Although the main approach of this course is hands-on practice, it is necessary to communicate among each other during the course. Most of the lectures were given by students in the class.

9.1.4.1 Introduction

- My View on Bioinformatics (Luo JC, CBI) [PDF]

9.1.4.2 UNIX and EMBOSS

- Unix – Carver T (EBI) [PDF] | Tian YL (CAAS06)F [PDF]
- EMBOSS – Fan L (CAAS07F) [PDF]

9.1.4.3 BLAST and Database Search

- BLAST – Luo JC (CBI) [PDF] | Tian YL (CAAS06F) [PDF] | Xie C (PKU08S1) [PDF] | Bian Y (PKU08S1) [PDF]
- Database search – Gao G (CBI) [PDF]
- Multiple sequence alignment – Liu K (PKU07S) [PDF]
- Scoring Matrix – Yuan YX (PKU07S) [PDF]

9.1.4.4 Resources

- PubMed – Huang LY (CAAS07F) [PDF] | Luo JC (CBI) [PDF]
- NCBI Databases – Li L (CAAS07F) [PDF]

- ExPASy – Gong JY (CAAS07F) [PDF]
- PDB – Gao HB (CAAS07F) [PDF]
- SRS – Huang BY (CAAS07F) [PDF]

9.1.4.5 Phylogeny

- MEGA – Hu YP (CAAS06F) [PDF]
- Phylogeny – Gao G (CBI) [PDF] | Huang BY (CAAS07F) [PDF] | Li Z (CBI) [PDF] | Shi XL (CBI) [PDF] | Zhu HW (CAAS07F) [PDF] | Li J (PKU08S) [PDF] | Yang Q (PKU08S) [PDF]

9.1.4.6 Molecular Modeling

- 3D modeling – Ye ZQ (CBI) [PDF] | Zhu HW (CAAS07F) [PDF] | Wang JL (CAAS08S) [PDF]
- 3D of Transcription Factors – Luo JC (CBI) [PDF]
- Swiss PDB Viewer – Li W (CAAS07F) [PDF]

9.1.4.7 Projects

- Hemoglobin – Li WQ (PKU07S) [PDF]
- MDR – Xu LM (PKU08S1) [PDF]
- CEA – Wang Q (CAAS07F) [PDF]
- SBP – Guo AY (CBI) [PDF]
- Rice – Zhu QH (CBI) [PDF]
- CA3 – Hou XH (CAAS07F) [PDF]
- PhyB – Wu FQ (CAAS07F) [PDF]
- P53 – Shen Y (CAAS07F) [PDF]
- Text Mining – Wang X (CAAS07F) [PDF]

All the PDF files of the above lectures can be downloaded freely for teaching. The copyright belongs to the original authors.

9.1.5 Exercises

As a hands-on practical course, we will introduce many exercises to be practiced during the course. This page collects a variety of exercises such as literature search, database query, and database search, sequence alignment, motif search, phylogenetic analysis, and molecular modeling.

Entrez – Literature search with PubMed and database query with the NCBI Entrez system.
ExPASy – Find protein sequences from the Swiss-Prot database with the ExPASy system.
SRS – Database query with the EBI Sequence Retrieval System (SRS).
Dotplot – DNA and protein sequence comparison using the dot plot approach.
Align – Pairwise and multiple sequence alignment using both local and global algorithms.
BLAST – Sequence similarity search against DNA and protein sequence databases.
DNA sequence analysis – Analysis of DNA sequences with several bioinformatics tools.
Protein sequence analysis – Analysis of protein sequences with several bioinformatics tools.
Motif – Identification of conserved sequence motifs and domains.
Phylogeny – Phylogenetic analysis and construction of phylogenetic tress with simple examples.
Modeling – Analysis of three-dimensional structure of proteins and prediction of protein structures.

9.2 Entrez

Literature search with PubMed and database query with the NCBI Entrez system.

9.2.1 PubMed Query

Find papers with the following keywords and compare the query results:

- "Hemoglobin," "human hemoglobin," "human hemoglobin AND structure," and "human hemoglobin AND function"
- "Hemoglobin [TI]," "human hemoglobin [AB]", "human hemoglobin [TI] AND structure [TIAB]," and "human hemoglobin [TI] AND structure [TIAB] AND Function [TIAB]"
- "Hemoglobin [TI] OR haemoglobin [TI]" and "hemoglobin [TI] OR haemoglobin [TI] AND structure [TIAB]"

Find papers with the following author names and compare the query results:

- "Sodmergen"
- "Danchin" and "Danchin A"
- "Li," "Li Y," and "Li YX"
- "Smith," "Smith T," and "Smith TT"

Find papers with the following query and compare the query results:

- "Rice," "rice [TI]," "rice [AU]," and "rice [AU] AND rice"
- "Rice," "Oryza sativa," and "rice OR Oryza sativa"
- "Luo J," "Luo JC," "Luo J[AU] AND bioinformatics," and "Luo J[AU] AND Peking University[AD]"
- "Luo J[AU] AND Peking University[AD] AND bioinformatics OR database OR rice"

9.2.2 Entrez Query

- Find the GenBank entry of the post-floral-specific gene (PPF-1) with the following query and compare query results: "Y12618," "PPF-1," and "post-floral-specific gene."
- Find the GenPept entry of the post-floral-specific protein (PPF-1) with the following query and compare query results: "Q9FY06," "PPF-1," and "post-floral-specific protein."
- Find the three-dimensional structure of the bar-headed goose hemoglobin.

9.2.3 My NCBI

- Register in My NCBI and make the following query: "bioinformatics [TI]."
- Save the above query and set the email delivery options.
- Configure the display options to show query results.

9.3 ExPASy

Find protein sequences from the Swiss-Prot database with the ExPASy system.

9.3.1 Swiss-Prot Query

- Find protein sequence of human hemoglobin alpha chain.
- Find protein sequence of mouse and rat hemoglobin alpha chains.
- Find protein sequence of human hemoglobin beta chain.
- Find protein sequence of mouse and rat hemoglobin beta chains.
- Find all entries of hemoglobin alpha chain in Swiss-Prot.
- Find all entries of hemoglobin beta chain in Swiss-Prot.

9.3.2 Explore the Swiss-Prot Entry HBA_HUMAN

- Retrieve the mRNA and coding sequence of human hemoglobin through cross link to GenBank.
- Find how many entries are deposited to Swiss-Prot through the Taxon cross link.
- Find the literatures related to crystal structure of human hemoglobin.
- Make a summary of the annotation of human hemoglobin based on the Comments of this entry.
- Make a summary of mutation of human hemoglobin based on the sequence features.
- Find out the alpha helices of human hemoglobin alpha chain.
- Find out the two Histidines which bind to the heme.
- Retrieve the FASTA sequence from this entry.

9.3.3 Database Query with the EBI SRS

SRS stands for sequence retrieval system. It is actually the main bioinformatics database query platform maintained by EBI and other bioinformatics centers around the world. SRS was originally developed by Etzold and Argos at the European Molecular Biology Laboratory (EMBL) in the early 1990s. It was moved to EBI and being continually developed by a group led by Etzold during the middle of the 1990s. In 1998, SRS became the main bioinformatics product of the biotechnology company LION Biosciences. Although it was a commercial software, SRS had been free for academic use until 2006 when it was acquired by BioWisdom.

9.3.3.1 Query

- Find human hemoglobin alpha chain sequence from Swiss-Prot with ID "HBA_HUMAN" or Accession "P69905," use Quick Search, Standard Query Form and Extended Query Form, compare the difference of search steps, and query results.
- Find human, mouse, and rat hemoglobin alpha chain sequences simultaneously from Swiss-Prot using ID "HBA_HUMAN," "HBA_MOUSE," and "HBA_RAT"; use Standard Query Form.
- Find all hemoglobin alpha chain sequence from Swiss-Prot with ID "HBA_."

9.3.3.2 Display and Save

- Display human hemoglobin alpha chain sequence ("HBA_HUMAN") with different sequence formats.

- Save all hemoglobin alpha chain sequences (HBA_) with FASTA format.
- Choose accession, sequence length, and species to display and save all hemoglobin alpha chain sequences (HBA_).

9.3.3.3 Dotplot

DNA and protein sequence comparison using the dot plot approach:

- Draw a dotplot for a test DNA sequence with a tandem repeat; use different word sizes to compare the results.
- Draw a dotplot for a test protein sequence (SENESCENSE); use different word sizes to compare the results.
- Draw a dotplot for the fugu cosmid sequence (AF164138).
- Retrieve similar regions from (AF164138) and draw a dotplot.
- Find similar domains from human and mouse carcinoembryonic antigens (CEAM5_HUMAN, CEAM1_MOUSE).
- Find out the sequence pattern of the Drosophila slit protein sequence (P24014) with the online Dotlet web server or the EMBOSS dotmatcher program.
- Find the zinc proteinase domain between human MS2 cell surface antigen (P78325) and adamalysin II (P34179) from Crotalus adamanteus (eastern diamondback rattlesnake) venom.
- Find the special sequence feature of serine-repeat antigen protein precursor (P13823) using dotplot.
- Find the special sequence feature of human zinc finger protein (Q9P255).
- Find the special sequence feature of human ubiquitin C (NP_066289) using dotplot.

9.4 Sequence Alignment

Pairwise and multiple sequence alignments using both local and global algorithms.

9.4.1 Pairwise Sequence Alignment

Use Needle and Water to align the following sequence pairs with different scoring matrices and gap penalties; compare the results:

- "AFATCAT" and "AFASTCAT"
- "THEFASTCATCATCHESAFATRAT" and "THEFATCATCATCHESADEADRAT"
- "AREALFRIENDISAFRIENDINNEED" and "AFRIENDINNEEDISAFRIENDINDEED"

Use Needle and Water to align the following sequence pairs with different scoring matrices and gap penalties; compare the results:

- Human hemoglobin alpha chain (Swiss-Prot Accession: P69905) and human hemoglobin beta chain (Swiss-Prot Accession: P68871)
- Human hemoglobin alpha chain (Swiss-Prot Accession: P69905) and yellow lupine leghemoglobin (Swiss-Prot Accession: P02240)
- Pisum sativum post-floral specific protein 1 (PPF-1, Swiss-Prot Accession: Q9FY06) and Arabidopsis inner membrane protein (ALBINO3, Swiss-Prot Accession: Q8LBP4)
- The coding sequence of PPF-1, GenBank Accession: Y12618 and ALBINO3, GenBank Accession: U89272
- The full length mRNA sequence of PPF-1, GenBank Accession: Y12618 and ALBINO3, GenBank Accession: U89272
- The rice histidine transporter (GenPept Accession: CAD89802) and the Arabidopsis lysine and histidine transporter (GenPet Accession: AAC49885)

Use Needle to align the following three sequences between each other; make a summary of your analysis results:

- The protein sequence of human, mouse, and rat hemoglobin alpha chains (Swiss-Prot entry name: HBA_HUMAN, HBA_MOUSE, HBA_RAT)
- The coding sequence of human, mouse, and rat hemoglobin alpha chains (GenBank Accession: V00493, V00714, M17083)

9.4.2 Multiple Sequence Alignment

- Make multiple sequence alignment for the protein sequence of hemoglobin alpha chain from 7 vertebrates [FASTA].
- Make multiple sequence alignment for the protein sequence of 12 human globins [FASTA].
- Make multiple sequence alignment for the protein sequence of 15 Arabidopsis SBP transcription factors [FASTA]; use different programs (ClustalW, T-Coffee, and DIALIGN) and compare the results.
- Make multiple sequence alignment for the 9 repeat sequences of human ubiquitin C protein (NP_066289).
- Make multiple sequence alignment for spider toxin peptides [FASTA]; use manual editing to improve the results.

9.4.3 BLAST

Sequence similarity search against DNA and protein sequence databases.

9.4.3.1 Online BLAST

- Search a virtual peptide sequence (ACDEFGHI) against the NCBI NR database.
- Search a virtual random peptide sequence (ADIMWQVRSFCYLGHTKEPN) against the NCBI NR database.
- Search a virtual DNA sequence (ACGTACGTACGTACGTACGT) against the NCBI NR database.
- Search the delta sleep-inducing peptide (Swiss-Prot Accession: P01158) against the NCBI NR database.
- Search PPF-1 protein sequence (Swiss-Prot Accession: Q9FY06) against the Swiss-Prot database.
- Search OsHT01 protein sequence (CAD89802) against the NCBI plants database.
- Search OsHT01 protein sequence (CAD89802) against the NCBI NR database.
- Search protein sequence (Q57997) against the NCBI NR database.
- Search cytokine induced protein sequence (NP_149073) against the NCBI NR database.
- Search olfactory receptor protein sequence (NP_001005182) against the NCBI NR database.

9.4.3.2 Local BLAST

- Construct a local BLAST database with hemoglobin alpha subunit protein sequences retrieved from Swiss-Prot and do BLAST search locally with a query sequence [207hba.fasta].
- Construct a local BLAST database with hemoglobin beta subunit protein sequences retrieved from Swiss-Prot and do BLAST search locally with a query sequence.
- Build local BLAST database for maize transcription factors [zmtf-mrna.fasta, zmtf-pep.fasta], and do BLAST search to find SBP TFs in maize TFs using seed sequence of the SPL3 DNA-binding domain [atsbpd3.fasta].

9.5 DNA Sequence Analysis

Analysis of DNA sequences with several bioinformatics tools.

9.5.1 Gene Structure Analysis and Prediction

- Draw the gene structure of the Drosophila melanogaster homolog of human Down syndrome cell adhesion molecule (DSCAM) (AF260530).
- Predict the potential genes from the fugu cosmid sequence [af164138.fasta].

9.5.2 Sequence Composition

- Find GC content of 2 pairs of rice sequence fragment; draw a plot [rice–pair1.fasta, rice-pair2.fasta].
- Find the CpG island of the fugu cosmid sequence [af164138.fasta].
- Calculate the codon usage of the coding sequence of fugu MDR3 [fugu-mdr3-cds.fasta].
- Calculate the GC content of coding and noncoding sequences of fugu cosmid [af164138-cds.fasta, af164138-ncs.fasta].

9.5.3 Secondary Structure

- Predict the tandem repeat of the human herpes virus 7 gene locus (HH7TETRA).
- Predict the secondary structure of human mitochondria tRNA-Leu (AB026838).

9.6 Protein Sequence Analysis

Analysis of protein sequences with several bioinformatics tools.

9.6.1 Primary Structure

- Use EMBOSS Pepstat to calculate amino acid composition of Pisum sativum post-floral specific protein 1 (PPF-1, Q9FY06).
- Use EMBOSS Pepinfo to create flowcharts of Pisum sativum post-floral specific protein 1 (PPF-1, Q9FY06).
- Use ExPASy Protscale to create flowcharts to display various properties of the Pisum sativum post-floral specific protein 1 (PPF-1, Q9FY06).

9.6.2 Secondary Structure

- Predict the secondary structure of human hemoglobin alpha chain (P69905) and beta chain (P68871).
- Predict the secondary structure of the N-terminal domain of mouse carcinoembryonic antigen (CEAM1_MOUSE).
- Predict the potential coiled-coil region of the yeast amino acid biosynthesis general control protein (GCN4_YEAST).

9.6.3 Transmembrane Helices

- Use EMBOSS TMAP and ExPASy TMHMM, TMPred, TopPred, and SOUSI to predict the transmembrane helices of the Pisum sativum post-floral specific protein 1 (PPF-1, Q9FY06).
- Use EMBOSS TMAP and ExPASy TMHMM, TMPred, TopPred, and SOUSI to predict the transmembrane helices of sheep ovine opsin (OPSD_SHEEP).
- Use EMBOSS TMAP and ExPASy TMHMM, TMPred, TopPred, and SOUSI to predict the transmembrane helices of GCR2 protein (NP_175700).

9.6.4 Helical Wheel

- Draw the four helical wheels for the transmembrane helices of the Pisum sativum post-floral specific protein 1 (PPF-1, Q9FY06).
- Draw the helical wheel for the last two helices of human hemoglobin alpha chain (P69905).

9.7 Motif Search

Identification of conserved sequence motifs and domains.

9.7.1 SMART Search

- Search SMART to find conserved domain for the post-floral-specific protein 1 (PPF-1, Swiss-Prot Accession: Q9FY06).
- Search SMART to find conserved domain for the RNA-binding protein AT1G60000.
- Search SMART to find conserved domain for the transcription factor SBP protein (At1g27370).

9.7.2 MEME Search

- Search MEME to find conserved domain for the protein sequence of 15 Arabidopsis SBP transcription factors [15atsbp.fasta].

9.7.3 HMM Search

- Use HMMER hmmbuild to build HMM model (atsbpd.hmm) for 15 SBP DNA-binding domains (15atsbpd.fasta).
- Use HMMER hmmcalibrate to adjust the above HMM model (atsbpd.hmm).
- Use HMMER hmmsearch using the HMM model (atsbpd.hmm) to identify SBP DNA-binding domains against maize transcription factors using the above HMM model.
- Build HMM model for 15 SBP DNA-binding domains (15atsbpd.fasta).

9.7.4 Sequence Logo

- Create a sequence logo for the SNPs of 20 SARS coronaviruses [20sars.fasta].
- Create a sequence logo for the DNA-binding domain of 15 Arabidopsis SBP proteins [15atsbpd.fasta].

9.8 Phylogeny

Phylogenetic analysis and construction of phylogenetic tress with simple examples.

9.8.1 Protein

- Construct phylogenetic tree with the maximum parsimony method for the hemoglobin alpha chain from 7 vertebrates [FASTA].
- Construct phylogenetic tree with the maximum parsimony method for the hemoglobin alpha chain using different datasets from 209 species [209hba.fasta] based on the taxonomy table [209HBA.PDF].
- Construct phylogenetic tree with the distance method for 12 human globulins [FASTA].
- Construct phylogenetic tree with the maximum parsimony method for a segment of lysine/histidine transporter from 10 plants [FASTA].
- Construct phylogenetic tree with both distance and maximum parsimony methods for the DNA-binding domain of 15 Arabidopsis SBP proteins [FASTA].

9.8.2 DNA

- Construct phylogenetic tree with distance and maximum parsimony methods for a set of 6 test DNA sequences [FASTA].
- Construct phylogenetic tree with the distance method for the SNPs of 20 SARS coronaviruses [FASTA].

9.9 Projects

As the name of this course implies, we focus on the application of bioinformatics tools to solve real problems in biological research. We chose several samples as working projects to learn how to find the literature, how to obtain sequence and structure data, how to do the analysis step by step, and how to make a summary from the analysis results. You are most encouraged to work on your own projects during the course:

PPF – sequence analysis of Pisum sativum post-floral specific gene
MDR – gene prediction and dotplot analysis of fugu multidrug resistance gene
BGH – sequence, structure, and function analysis of the bar-headed goose hemoglobin
CEA – protein engineering of carcinoembryonic antigen
AFP – structure comparison of spider toxins and prediction of antifungal peptide
HMT – sequence and structure comparison of human metallothioneins
SBP – systematic analysis of the Arabidopsis transcription factor family of Squamosa-promoter binding protein

9.9.1 Sequence, Structure, and Function Analysis of the Bar-Headed Goose Hemoglobin

Hemoglobin is one of the most well-studied proteins in the last century. The sequence, structure, and function of several vertebrates have been investigated during the past 50 years. More than 200 hundreds of hemoglobin protein sequences have been deposited into the Swiss-Prot database. Three-dimensional structure wild type and mutants from dozens of species have been solved. This provides us a good opportunity to study the relationship among sequence, structure, and function of hemoglobins.

Bar-headed goose is a special species of migration birds. They live in the Qinghai Lake during summer time and fly to India all the way along over the Tibetan plateau in autumn and come back in spring. Interestingly, a close relative of bar-headed goose, the graylag goose, lives in the low land of India all year round and do not migrate. Sequence alignment of bar-headed goose hemoglobin with that of graylag goose shows that there are only 4 substitutions. The Pro 119 in the alpha subunit

of graylag goose has been changed to Ala in bar-headed goose. This residue is located in the surface of the alpha/beta interface. In 1983, Perutz proposed that this substitution reduces the contact between the alpha and beta subunits and increases the oxygen affinity, due to the relation of the tension status in the deoxy form [1].

During the past decade, a research group at Peking University has solved the crystal structure of both deoxy and oxy forms of bar-headed goose as well as the oxy form of the graylag goose hemoglobin [2–4]. We will use this example to learn how to analyze the sequence and structure of hemoglobin molecules.

9.9.2 Exercises

- Find literature on bar-headed goose hemoglobin from PubMed.
- Find protein sequences of bar-headed goose and graylag goose hemoglobins from Swiss-Prot; make sequence alignment to compare the differences between these two molecules.
- Find protein sequences which share 90 % similarity with bar-headed hemoglobin alpha chain, make multiple sequence alignment, and construct a phylogenetic tree for the above protein sequences using the maximum parsimony method.
- Retrieve three-dimensional structure oxy (1A4F) and deoxy (1HV4) forms of bar-headed goose hemoglobin; compare the difference of the heme molecule with Swiss PDB Viewer.
- Make superimposition for the alpha and beta subunit of bar-headed goose (1A4F) and greylag goose (1FAW) hemoglobins, find out the differences of the substitution site (Pro119-Ala119), and measure the contact distance between the side chain of this residues and the side chain of Ile55 of the beta subunit.
- Make a summary about the sequence, structure, and function of bar-headed goose hemoglobin.

9.10 Literature

There are many online materials dedicated to bioinformatics education including courses, tutorials, and documents. You are most encouraged to get access to these self-educational websites during the course and for your further study.

9.10.1 Courses and Tutorials

- 2CAN – EBI bioinformatics support portal which provides short and concise introductions to basic concepts in molecular and cell biology and bioinformatics
- EMBER – an online practical course with multiple choice quiz designed and maintained by Manchester University, UK (free registration needed)

- Science Primer – NCBI science primer for various topics including bioinformatics, genome mapping, and molecular biology
- SADR – an online course for Sequence Analysis with Distributed Resources, Bielefeld University, Germany
- SWISS-MODEL – an online course for principles of protein structure, comparative protein modeling, and visualization maintained by Nicolas Guex and Manuel C. Peitsch at Glaxo Wellcome
- Swiss-PDB Viewer – an extensive tutorial for the molecular visualization and modeling program Swiss PDB Viewer, created and maintained by Gale Rhodes at the University of Southern Maine

9.10.2 Scientific Stories

- Molecular of the Month – A website which presents short accounts on selected molecules from the Protein Data Bank. It was created in January 2000 and is being maintained by David S. Goodsell at Scripps.
- Protein Spotlight – A website which tells short stories on protein molecules as well as the scientific research and the scientists behind these interesting stories. It was started in January 2000 and is being maintained by Vivienne B. Gerritsen at Swiss Institute of Bioinformatics.

9.10.3 Free Journals and Books

- NCBI Bookshelf – a growing collection of free online biomedical books that can be searched directly through the NCBI Entrez system.
- PubMed Central – the US National Institutes of Health (NIH) free digital archive of biomedical and life sciences journals.
- PubMed Central China mirror – the PubMed Central China mirror maintained by the Center for Bioinformatics, Peking University.
- BioMed Central – the website of more than 180 open access biomedical journals freely available, started by a UK publisher in London since 2001.
- PLOS – the website of several significant open access biological and medical journals freely available, started by the Public Library of Science, a nonprofit organization composed of many famous scientists.
- HighWire – a website maintained by the Stanford University Libraries. It gives a list of biomedical journals which provide either immediate or 6/12/18/24 months delay of free online full-text articles.
- Amedeo – a website maintained by Bernd Sebastian Kamps in Europe. It provides extensive links to various biomedical journals and books freely available.
- AnaTax – an online book chapter of the anatomy and taxonomy written in 1981 and being updated by Jane Richardson at Duke University.

9.11 Bioinformatics Databases

There are huge amount of online bioinformatics databases available on the Internet. The databases listed in this page are extensively accessed during this course.

9.11.1 List of Databases

- NAR databases – the most extensive list of biological databases being maintained by the international journal *Nucleic Acids Research* which publishes a special issue for molecular biology databases in the first issue of each year since 1996. All these database papers can be accessed freely. You may find links to the website of the databases described in the chapter.
- NCBI databases – the molecular databases maintained by NCBI. A Flash flowchart for 24 databases connected by lines shows the relationships and internal links among all these databases. These databases are divided into 6 major groups: nucleotide, protein, structure, taxonomy, genome, and expression. It also provides links to the individual database description page.
- EBI databases – the main portal to all EBI databases divided in several groups, such as literature, microarray, nucleotide, protein, structure, pathway, and ontology. Links to database query and retrieval systems can be found in this portal.

9.11.2 Database Query Systems

- NCBI Entrez – the unique interface to search all NCBI databases. Query results are displayed with entry numbers along the database names.
- EBI SRS – the database query system maintained by EBI. SRS stands for sequence retrieval system which was originally developed by Thure Etzold at the European Molecular Biology Laboratory in the early 1990s and moved to EBI. It was originally an open system and installed in a dozen of institutions with different databases. In the late 1990s, SRS became a commercial package but still free to academic use. In 2006, SRS was acquired by BioWisdom, a software company based in Cambridge, UK.
- EMBL SRS – the database query system (ver. 8.2) maintained by EMBL, at Germany.
- DKFZ SRS – the database query system (ver. 8.2) maintained by the German Cancer Research Center.
- Columbia SRS – the database query system (ver. 8.2) maintained by Columbia University, USA.

- CMBI MRS – the open-source database query system developed by Marteen Hekkelman at the Center for Molecular and Biomolecular Information (CMBI), the Netherlands.
- CBI MRS – the MRS installed at the Center for Bioinformatics, Peking University.

9.11.3 Genome Databases

Due to the rapid progress of DNA sequencing technology, hundreds of organisms have been sequenced at the genome scale. Genome databases and related analysis platforms can be accessed on the Internet. We select some of them for our teaching purpose of this course.

9.11.3.1 List of Genome Databases

- GOLD – Genomes Online Database, a comprehensive information resource for complete and ongoing genome sequencing projects with flowcharts and tables of statistical data. Created and maintained by
- Karyn's Genome – a collection of sequenced genomes with brief description and references for each genome. The direct links to the sequence data in EMBL or annotations in ENSEMBL make it very convenient for the user community.
- CropNet – the website of the UK Crop Plant Bioinformatics Network to the development, management, and distribution of information relating to comparative mapping and genome research in crop plants.

9.11.3.2 Genome Browsers and Analysis Platforms

- NCBI Genome – the entry portal to various NCBI genomic biology tools and resources, including the Map Viewer, the Genome Project Database, and the Plant Genomes Central
- GoldenPath – the genome browser website containing the reference sequence and working draft assemblies for a large collection of genomes at the University of California at Santa Cruz (UCSC)
- ENSEMBL – the web server of the European eukaryotic genome resource developed by EBI and the Sanger Institute [PDF]
- VISTA – a comprehensive suite of programs and databases for comparative analysis of genomic sequences
- TIGR Plant Genomics – TIGR plant genome databases and tools
- TIGR Gene Indices – TIGR gene indices maintained at Harvard

9.11.3.3 Genome Database of Model Organisms

- Gramene – a curated open-source data resource for comparative genome analysis in the grasses including rice, maize, wheat, barley, and sorghum, as well as other plants including arabidopsis, poplar, and grape. Cross-species homology relationships can be found using information derived from genomic and EST sequencing, protein structure and function analysis, genetic and physical mapping, interpretation of biochemical pathways, gene and QTL localization, and descriptions of phenotypic characters and mutations.
- TAIR – the Arabidopsis information resource maintained by Stanford University. It includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.
- AtENSEMBL – a genome browser for the commonly studied plant model organism Arabidopsis thaliana.
- Oryzabase – a comprehensive rice science database maintained by the National Institute of Genetics, Japan. It contains genetic resource stock information, gene dictionary, chromosome maps, mutant images, and fundamental knowledge of rice science.
- FlyBase – a comprehensive database of Drosophila genes and genomes maintained by Indiana University.
- CyanoBase – the genome database for cyanobacteria developed by Kazusa Institute, Japan.

9.11.4 Sequence Databases

DNA and protein sequence databases are the fundamental resources for bioinformatics research, development, and application.

9.11.4.1 DNA Sequence Databases

- GenBank – the web portal to the NIH genetic sequence database maintained by NCBI, also a part of the International Nucleotide Database Collaboration. Literature citation, release notes, and an example record can be found in this page.
- EMBL – the web portal to EMBL nucleotide sequence database maintained by EBI, also a part of the International Nucleotide Database Collaboration. Various documentations such as release notes, database statistics, user guide, feature table definition and sample entry, and FAQs are provided.

- RefSeq – the Reference Sequence collection constructed by NCBI to provide a comprehensive, integrated, non-redundant set of DNA and RNA sequences and protein products. It provides a stable reference for genome annotation, gene identification and characterization, mutation and polymorphism analysis, expression studies, and comparative analyses.
- UniGene – an Organized View of the Transcriptome created by NCBI. Each UniGene entry is a set of transcript sequences that appear to come from the same transcription locus, together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location.
- dbSNP – the database of single nucleotide polymorphism maintained by NCBI.
- EMBLCDS – a database of nucleotide sequences of the coding sequence from EMBL.

9.11.4.2 Protein Sequence Databases

- Swiss-Prot – the entry site for the well-annotated protein sequence knowledge database maintained by SIB at Geneva, Switzerland. A list of references, a comprehensive user manual, and database statistics with tables and flowcharts are provided.
- UniProt – the main website for international protein sequence database which consists of the protein knowledgebase (UniProtKB), the sequence clusters (UniRef), and the sequence archive (UniParc).
- HPI – the Human Proteomics Initiative, an EBI project to annotate all known human sequences according to the quality standards of UniProtKB/Swiss-Prot. It provides for each known protein a wealth of information that includes the description of its function, its domain structure, subcellular location, posttranslational modifications, variants, and similarities to other proteins.
- IPI – the website of the International Protein Index database which provides a top level guide to the main databases that describe the proteomes of higher eukaryotic organisms.

9.11.5 Protein Domain, Family, and Function Databases

Protein molecules play functions in living organisms. They are usually classified into families based on the different functions they play. Proteins in the same family or subfamily often have conserved sequence motifs or fingerprints as well as ungapped blocks or functional domains. With the great amount of available sequence data, secondary databases of protein molecules have been constructed. We select some of them for this course.

9.11.5.1 Protein Domain Databases

- Prosite – a database of protein domains, families, and functional sites, created and maintained by the Swiss Institute of Bioinformatics
- PRINTS – a database of protein fingerprints consisting of conserved motifs within a protein family, created and maintained by Manchester University, UK
- BLOCKS – a database of multiple aligned ungapped segments corresponding to the most highly conserved regions of proteins, created and maintained by the Fred Hutchinson Cancer Research Center, USA
- CDD – a database of conserved protein domains created and maintained by the NCBI structure group
- ProDom – a database of comprehensive set of protein domain families automatically generated from the Swiss-Prot and TrEMBL sequence databases, developed and maintained by the University Claude Bernard, France

9.11.5.2 Protein Family Databases

- Pfam – a database of protein families represented by multiple sequence alignments and hidden Markov models, constructed and maintained by the Sanger Institute, UK

9.11.5.3 Protein Function Databases

- IMGT – the international immunogenetics information system, a high-quality integrated knowledge resource of the immune system of human and other vertebrate species, created and maintained by the University of Montpellier, France
- HPA – a website for the human protein atlas which shows expression and localization of proteins in a large variety of normal human tissues, cancer cells, and cell lines with the aid of immunohistochemistry images, developed and maintained by Proteome Resource Center, Sweden

9.11.6 Structure Databases

The central point of the protein structure database is Protein Data Bank (PDB) which was started in the late 1970s at the US Brookhaven National Laboratory. In 1999, the Research Collaboratory for Structural Bioinformatics (RSCB) was formed to manage the PDB. In 2003, an international collaboration among RSCB, MSD-EBI at Europe, and PDBj in Japan was initiated to form wwPDB, and the Magnetic

Resonance Data Bank (BMRB) joined wwPDN in 2006. Although RSCB is the main entry point to access macromolecular structures, we may also find protein structures through the NCBI Entrez system and the EBI MSDLite server.

9.11.6.1 Main Portals for the Protein Structures Database

- RSCB – the main repository of macromolecular structures maintained by the Research Collaboration for Structural Bioinformatics
- MMDB – the macromolecular database maintained by NCBI
- MSD – the entry point for the EBI macromolecular structure database
- MSDLite – the EBI web server providing simple search of protein structures
- PDBSUM – the EBI web server which provides overview, schematic diagrams, and interactions of structure
- BMRB – the biological magnetic resonance data bank maintained at University of Wisconsin-Madison
- ModBase – the database of comparative protein structure models developed and maintained at University of California, San Francisco

9.11.6.2 Classification of the Protein Structures

- SCOP – the database of Structure Classification of Proteins developed and maintained by Cambridge University
- CATH – the database of Calcification, Architecture, Topology, and Homologous superfamily developed and maintained by the University College London

9.11.6.3 Visualization of Protein Structures

- JenaLib – the Jena Library of Biological Macromolecules which provides information on macromolecular structures with an emphasis on visualization and analysis

9.12 Bioinformatics Tools

There are many bioinformatics tools over the Internet. Thanks to the great contribution of the scientific community of bioinformatics research and development. They make the bioinformatics programs and packages freely available to the end user biologists. The web-based tools can be accessed through the Internet using the web browsers such as Firefox and Internet Explorer. Users may also download and install some of the packages on a local machine to run the programs. We use both web-based platforms as well as command line tools integrated in a Linux-based

bioinformatics environment Bioland developed locally at our center. The three major international bioinformatics centers, NCBI, EBI, and ExPASy, develop, collect, and maintain hundreds of bioinformatics tools. They also provide online service for most of these web-based tools.

9.12.1 List of Bioinformatics Tools at International Bioinformatics Centers

- ExPASy tools – a comprehensive list of online web-based bioinformatics tools provided by ExPASy and worldwide
- EBI tools – the entry page for the EBI bioinformatics tools
- NCBI tools – the entry page for the NCBI bioinformatics tools

9.12.2 Web-Based Bioinformatics Platforms

- WebLab – the comprehensive and user-friendly bioinformatics platform developed by the Center for Bioinformatics, Peking University. WebLab provides user spaces to store and manage input data and analysis results as well as literature references. The analysis protocols and macros allow users to process the job in a batch mode.
- EMBOSS explorer – the web interface for the EMBOSS package, maintained by Ryan Golhar at the University of Medicine and Dentistry of New Jersey.
- EMBOSS – the web interface for the EMBOSS package, maintained by the University of Singapore.
- SRS Tools – the EBI SRS database query server integrates several analysis packages such as EMBOSS and HMMER which can be launched directly with retrieved data from the SRS server or with external data provided by users.

9.12.3 Bioinformatics Packages to be Downloaded and Installed Locally

- EMBOSS – the main portal for the open-source bioinformatics project EMBOSS (European Molecular Biology Open Software Suite) headed by Peter Rice and Alan Bleasby at EBI. EMBOSS is a comprehensive package with some 200 individual programs for DNA and protein sequence analysis.
- PISE – the main page to learn and download the PISE software designed by the Pasteur Institute, France. It can generate the web interface for the programs of the EMBOSS package.

- wEMBOSS – the entry page for a simple description and download of the EMBOSS graphics user interface.
- wEMBOSS – the web page for the HMMER package which uses profile hidden Markov models to detect conserved domains in protein sequences, developed and maintained by Sean Eddy at Howard Hughes Medical Institute.

This page lists the most comprehensive packages such as EMBOSS. For other packages and programs, please find them in the individual pages list in the Tools menu.

9.13 Sequence Analysis

DNA and protein sequence analysis is the most fundamental approach in bioinformatics.

9.13.1 Dotplot

- Dotlet – the Java-based web server for the comparison of DNA and protein sequences using the dot plot approach, maintained by the Swiss Institute of Bioinformatics.

9.13.2 Pairwise Sequence Alignment

- Align – the web server for the pairwise sequence alignment, maintained by EBI. Either global (Needle) or local (Water) alignment can be performed.
- BLAST 2 – the web server for the alignment of two sequences using BLAST maintained at NCBI.

9.13.3 Multiple Sequence Alignment

- ClustalW – the web server for the global multiple sequence alignment program maintained at EBI. It uses a new version of ClustalW 2.0.
- Muscle – the web server for the multiple sequence comparison by log-expectation, maintained at EBI. It claimed to achieve better accuracy with higher speed than ClustalW depending on the chosen options.
- MAFAT – the web server of high speed multiple sequence alignment using fast Fourier transformation, maintained at EBI.
- KALIGN – the web server of fast and accurate multiple sequence alignment, maintained at EBI.

- T-Coffee – the web server of several tools for computing, evaluating, and manipulating multiple alignments of DNA and protein sequences and structures, developed and maintained by Cedric Notredame at the Center for Genomic Regulation, Spain.
- DIALIGN – the web server for the multiple sequence alignment, developed and maintained by Burkhard Morgenstern at Bielefeld University, Germany. It uses a local alignment approach to compare a whole segment of sequences without gap penalty.

9.13.4 Motif Finding

- SMART – the web server for motif discovery and search, developed, and maintained by the University of California at San Diego. It provides
- MEME – the web server for motif discovery and search, developed and maintained by the University of California at San Diego. It provides
- TMHMM – the web server for the prediction of transmembrane helices in proteins, developed and maintained by Denmark Technical University.

9.13.5 Gene Identification

- GENSCAN – the web server for the identification of complete gene structures in genomic DNA, maintained by MIT
- GenID – the web server for the prediction of genes in anonymous genomic sequences, developed and maintained by the University of Pompeu Fabra, Spain
- HMMGene – the web server for the prediction of vertebrate and C elegans genes, developed and maintained by the Denmark Technical University
- SoftBerry – the web server for gene prediction, limited free use for academic users

9.13.6 Sequence Logo

- WebLogo – the web server for the generation of sequence logos, developed and maintained by the University of California at Berkeley.

9.13.7 RNA Secondary Structure Prediction

- MFOLD – the web server for the prediction of RNA secondary structure, developed and maintained by Michael Zuker at Rensselaer Polytechnic Institute.

9.14 Database Search

Sequence similarity search against DNA or protein sequence database is one of the most extensively used approaches in the application of bioinformatics to molecular biology and genome research. Results of database search can deduce biological function for the newly identified sequence or to infer evolutionarily relationship among the query sequence and a group of matched subject sequences. The most famous program for database search is BLAST – the Basic Local Alignment Search Tool. Thanks to the BLAST team at NCBI who continuously develop this package and make it freely available for the community. Here, we introduce several BLAST web servers which are mostly accessed as well as the FASTA, BLAT, and MPSearch servers.

9.14.1 BLAST Search

- NCBI BLAST – the central point of the NCBI BLAST server which provides whole functionality of DNA and protein sequence database search programs including PSI-BLAST and PHI-BLAST and the whole list of databases in different scales and different genomes.
- EBI BLAST – the entry page of the EBI BLAST facility which provides both NCBI BLAST and WU-BLAST programs as well as a set of specialized BLAST programs to search the alternative splicing database, the cloning vector database, and the parasite database.
- WU-BLAST – the entry page of the WU-BLAST maintained by Warren Gish at Washington University. It provides several links to other BLAST servers worldwide.
- Sanger BLAST – the entry page of the sequence projects BLAST search services at the Sanger Institute, featured with the special genome databases generated by the completed or ongoing sequencing projects.

9.14.2 Other Database Search

- FASTA – the original website for the FASTA set programs maintained by William Pearson at the University of Virginia. It provides extensive documents and help materials as well.
- EBI FASTA – the entry page of EBI FASTA services. FASTA has a higher sensitivity in some cases comparing with BLAST with a slightly lower speed. It is specific to identify low similarity long regions for highly diverged sequences.

- MPSrch – the EBI MPSrch server which uses the Smith-Waterman algorithm to obtain the optimal sequence alignment results.
- GoldenPath BLAT – the web server for the genome database search, integrated in the GoldenPath platform.
- ENSEMBL BLAT – the web server for the genome database search, integrated in the ENSEMBL platform.

9.15 Molecular Modeling

Molecular modeling is one of the important disciplines of bioinformatics. The latest development of both hardware and software makes it possible for molecular visualization which is fundamental for molecular modeling. Currently, there are quite few plug-ins for the real-time display and manipulation of three-dimensional structures with Internet browsers such as Jmol and WebMol provided in the PDB web server. You may also install stand-alone tools on your PC such as Swiss PDB Viewer and PyMOL which have more functionalities. On the other hand, homology-based protein modeling web servers may help you to predict the three-dimensional structure of your protein based on sequence similarity between your protein and the templates with known 3D structures.

9.15.1 Visualization and Modeling Tools

- Swiss PDB Viewer – a comprehensive molecular visualization and modeling package developed by Nicolas Guex at GlaxoSmithKline. It provides a graphical interface and a user-friendly control panel for users to analyze several proteins at the same time. There are lots of useful functions such as superimposition, mutation, and energy minimization for simple molecular simulations.
- PyMOL – the home page of the molecular visualization system written in Python. It is a user-sponsored open-source software maintained by DeLano Scientific LLC.
- Cn3D – the web page of the molecular visualization program Cn3D developed and distributed by NCBI. It can be installed on your PC to display the three-dimensional structures obtained from the NCBI MMDB database.
- Kinemage – the home page of the protein visualization package Kinemage developed and maintained at the laboratory of Jane Richardson and David Richardson, Duke University.
- RasMol – the home page of molecular visualization freeware RasMol, maintained by the University of Massachusetts, Amherst.

9.15.2 Protein Modeling Web Servers

- Swiss Model – the home page of the automated comparative protein modeling server, hosted by the University of Basel and the Swiss Institute of Bioinformatics. Extensive documentation and help materials are provided.
- 3D Jigsaw – the web server to build three-dimensional models for protein molecules based on homologues of known structure, developed and maintained by the Cancer Research UK.
- CASP – the entry page for the international protein structure prediction center, hosted at University of California at Davies. It provides the means of objective testing of evaluation of different methods in protein structure prediction.

9.16 Phylogenetic Analysis and Tree Construction

There are more than 300 hundreds phylogeny programs available on the Internet. Most of them can be downloaded and installed freely on your own machine. Due to the great need of computing power, it is difficult to maintain online phylogenetic analysis web servers. The best way to do phylogenetic analysis is to use command line for the PHYLIP programs integrated in EMBOSS or install MEGA on your PC Windows.

9.16.1 List of Phylogeny Programs

- Phylogeny software – the whole list of phylogeny programs collected and classified in groups by Joe Felsenstein.

9.16.2 Online Phylogeny Servers

- WebLab Protocols – the WebLab platform we develop and maintain has integrated the PHYLIP package. The protocols and macros for both Neighbor Joining and maximum parsimony methods are extremely useful for biologists to construct phylogeny trees with well-defined data sets.
- NUS EMBOSS interface – the web interface of the PHYLIP programs integrated in the EMBOSS package, maintained by the University of Singapore.
- EBC interface – the web interface of some PHYLIP programs, maintained by Uppsala University, Sweden.

9.16.3 Phylogeny Programs

- PHYLIP – the website for the comprehensive package Phylogeny Inference Package (PHYLIP) created and maintained by Joe Felsenstein at the University of Washington. This package can be downloaded and installed on Linux, Windows, and Mac freely.
- TREE-PUZZLE – the web page for the phylogeny program which uses the maximum likelihood method to analogize nucleotide and amino acid sequences as well as other two-state data.
- PAML – the website for the Phylogenetic Analysis by Maximum Likelihood package developed and maintained by Ziheng Yang, at University College, London.
- MEGA – the website for the Molecular Evolutionary Genetics Analysis package developed and maintained by Masatoshi Nei and his colleagues. It was originally designed for the Windows platform with a graphics interface and uses the distance method to construct phylogenetic trees [PDF].

9.16.4 Display of Phylogenetic Trees

iTOL – the website of the Interactive Tree of Life for the display and manipulation of phylogenetic trees, developed and maintained by the European Molecular Biology Laboratory.

References

1. Perutz MF (1983) Species adaptation in a protein molecule. Mol Biol Evol 1(1):1–28
2. Zhang J, Hua Z, Tame JR, Lu G, Zhang R, Gu X (1996) The crystal structure of a high oxygen affinity species of haemoglobin. J Mol Biol 255(3):484–493
3. Liang YH, Liu XZ, Liu SH, Lu GY (2001) The structure of greylag goose oxy haemoglobin: the roles of four mutations compared with bar-headed goose haemoglobin. Acta Crystallogr D Biol Crystallogr 57(Pt 12):1850–1856
4. Liang Y, Hua Z, Liang X, Xu Q, Lu G (2001) The crystal structure of bar-headed goose hemoglobin in deoxy form: the allosteric mechanism of a hemoglobin species with high oxygen affinity. J Mol Biol 313(1):123–137