Appendix

APPENDIX ONE

Practical Exercises

Note: all exercises were originally designed for use on a UNIX workstation. However, with slight modifications, they can be used on any other operating systems with Internet access.

EXERCISE 1. DATABASE SEARCHES

In this exercise, you will learn how to use several biological databases to retrieve information according to certan criteria. After learning the basic search techniques, you will be given a number of problems and asked to provide answers from the databases.

1. Use a web browser to retrieve a protein sequence of lambda repressor from SWISS-PROT (http://us.expasy.org/sprot/). Choose "Full text search in Swiss-Prot and TrEMBL." In the following page, Enter "lambda repressor" (space is considered as logical operator AND) as keywords in the query window. Select "Search in Swiss-Prot only." Click on the "submit" button. Note the search result contains hypertext links taking you to references that are cited or to other related information. Spend a little time studying the annotations.
2. In the same database, search more sequences for "human MAP kinase inhibitor," "human catalase," "synechocystis cytochrome P450," "coli DNA polymerase," "HIV CCR5 receptor," and "*Cholera* dehydrogenase." Record your findings and study the annotations.
3. Go to the SRS server (http://srs6.ebi.ac.uk/) and find human genes that are larger than 200 kilobase pairs and also have poly-A signals. Click on the "Library Page" button. Select "EMBL" in the "Nucleotide sequence databases" section. Choose the "Extended" query form on the left of the page. In the following page, Select human ("hum") at the "Division" section. Enter "200000" in the "SeqLength >=" field. Enter "polya_signal" in the "AllText" field. Press the "Search" button. How many hits do you get?
4. Use your knowledge and creativity to do the following SRS exercises.
   1) Find protein sequences from *Rhizobium* submitted by Ausubel between 1991 and 2001 in the UniProt/Swiss-Prot database (hint: the date

expression can be 1-Jan-1991 and 31-Dec-2001). Study the annotations of the sequences.

2) Find full-length protein sequences of mammalian tyrosine phosphatase excluding partial or fragment sequences in the UniProt/SwissProt database (hint: the taxonic group of mammals is *mammalia*). Once you get the query result, do a Clustal multiple alignment on the first five sequences from the search result.

5. Go to the web page of NCBI Entrez (http://www.ncbi.nlm.nih.gov/) and use the advanced search options to find protein sequences for human kinase modified or added in the last 30 days in GenBank. In the Entrez "Protein" database, enter "human[ORGN] kinase", and then select "last 30 days" in the "Modification Date" field of the "Limits" section. Select "Only from" "GenBank" as database. Finally, select " Go."
6. Using Entrez, search DNA sequences for mouse fas antigen with annotated exons or introns. (Do not forget to deselect "Limits" from the above exercise.) In Entrez, select the Nucleotide database. Type mouse[ORGN] AND fas AND (exons OR introns). Click "Go."
7. For the following exercises involving the NCBI databases, design search strategies to find answers (you will need to decide which database to use first).
   1) Find gene sequences for formate dehydrogenase from *Methanobacterium*.
   2) Find gene sequences for DNA binding proteins in *Methanobacterium*.
   3) Find all human nucleotide sequences with D-loop annotations.
   4) Find protein sequences of maltoporin in Gram-negative bacteria (hint: use logic operator NOT. Gram-positive bacteria belong to Firmicutes).
   5) Find protein structures related to *Rhizobium* nodulation.
   6) Find review papers related to protein electrostatic potentials by Honig published since 1990.
   7) Find the number of exons and introns in *Arabidopsis* phytochrome A (*phyA*) gene (hint: use [GENE] to restrict search).
   8) Find two upstream neighboring genes for the hypoxanthine phosphoribosyl transferase (HPRT) gene in the *E. coli* K12 genome.
   9) Find neurologic symptoms for the human Lesch–Nyhan syndrome. What is the chromosomal location of the key gene linked to the disease? What are its two upstream neighboring genes?
   10) Find information on human gene therapy of atherosclerosis from NCBI online books.
   11) Find the number of papers that Dr. Palmer from Indiana University has published on the subject of lateral gene transfer in the past ten years.

EXERCISE 2. DATABASE SIMILARITY SEARCHES AND PAIRWISE SEQUENCE ALIGNMENT

Database Searching

In this exercise, you will learn about database sequence similarity search tools through an example: flavocytochrome b2 (PDB code 1fcb). This enzyme has been shown to be very similar to a phosphoribosylanthranilate isomerase (PDB code 1pii) by detailed three-dimensional structural analysis (Tang et al. 2003. *J. Mol. Biol.* 334:1043–62). This similarity may not be detectable by traditional similarity searches. Perform the following exercise to test the capability of various sequence searching methods to see which method has the highest sensitivity to detect the distant homologous relationship.

1. Obtain the yeast flavocytochrome b2 protein sequence from NCBI Entrez (accession number NP_013658). This is done by choosing "FASTA" in the format pull-down menu and clicking on the "Display" button. Copy the sequence into clipboard.
2. Perform a protein BLAST search (select Protein-protein BLAST at www.ncbi.nlm.nih.gov/blast/). Paste the sequence into the BLASTP query box. Choose pdb as database (this will reduce the search time). Leave all other settings as default. Click on the "BLAST!" button. To get the search result, click on the "Format!" button in the following page. Summarize the number of hits, highest and lowest bit scores in a table.
3. Change the *E*-value to 0.01 and change the word size from 3 to 2, and do the search again. Do you see any difference in the number of hits? Can you find 1pii in the search result?
4. Reset the *E*-value to 10. Change the substitution matrix from BLOSUM62 to BLOSUM45. Compare the search results again. What is your conclusion in terms of selectivity and sensitivity of your searches? Record the number of hits, and the highest and lowest scores in a table.
5. Reset the substitution matrix to BLOSUM62, run the same search with and without the low-complexity filter on. Compare the results.
6. Run the same search using FASTA (www.ebi.ac.uk/fasta33/). Choose pdb as database and leave other parameters as default. Compare the results with those from BLAST.
7. Run an exhaustive search using ScanPS (www.ebi.ac.uk/scanps/) using the default setting. This may take a few minutes. Compare results with BLAST and FASTA. Can you find 1pii in the result page?
8. Go back to the NCBI BLAST homepage, run PSI-BLAST of the above protein sequence by selecting the subprogram "PHI- and PSI-BLAST" (PHI-BLAST is pattern matching). Paste the sequence in the query box and choose pdb as database. Select "BLAST!" (for a regular query sequence, PSI-BLAST is automatically invoked). Click on "Format!" in the next page. The results will be

returned in a few minutes. Notice the symbols (New or green circle) in front of each hit.

9. By default, the hits with *E*-values below 0.005 should be selected for use in multiple sequence alignment and profile building. Click on the "Run PSI-Blast iteration 2" button. This refreshes the previous query page. Click the "Format!" button to retrieve results.
10. In the results page, notice the new hits generated from the second iteration. Perform another round of PSI-BLAST search. Record the number of hits and try to find 1pii in the result page.
11. Finally, do the same search using a hidden Markov model based approach. Access the HHPRED program (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred) and paste the same query sequence in the query window. Click the "Submit job" button.
12. The search may take a few minutes. When the results are returned, can you find 1pii in the search output?
13. Compare the final results with those from other methods. What is your conclusion regarding the ability of different programs to find remote homologs?

Pairwise Sequence Alignment

1. In the NCBI database, retrieve the protein sequences for mouse hypoxanthine phosphoribosyl transferase (HPRT) and the same enzyme from *E. coli* in FASTA format.
2. Perform a dot matrix alignment for the two sequences using Dothelix (www.genebee.msu.su/services/dhm/advanced.html). Paste both sequences in the query window and click on the "Run Query" button. The results are returned in the next page. Click on the diagonals on the graphic output to see the actual alignment.
3. Perform a local alignment of the two sequences using the dynamic programming based LALIGN program (www.ch.embnet.org/software/LALIGN_form.html). Make sure the two sequences are pasted separately in two different windows. Save the results in a scratch file.
4. Perform a global alignment using the same program by selecting the dial for "global." Save the results and compare with those from the local alignment.
5. Change the default gap penalty from "−14/−4" to "−4/−1". Run the local alignment and compare with previous results.
6. Do a pairwise alignment using BLAST (in the BLAST homepage, select the bl2seq program). Compare results with the previous methods.
7. Do another alignment with an exhaustive alignment program SSEARCH (http://pir.georgetown.edu/pirwww/search/pairwise.html). Compare the results.
8. Run a PRSS test to check whether there is any statistically significant similarity between the two sequences. Point your browser to the PRSS web page (http://fasta.bioch.virginia.edu/fasta/prss.htm). Paste the sequences in the

FASTA format in the two different windows. Use 1,000 shuffles and leave everything else as default. Click on the "Compare Sequence" button. Study the output and try to find the critical statistical parameters.

EXERCISE 3. MULTIPLE SEQUENCE ALIGNMENT AND MOTIF DETECTION

Multiple Sequence Alignment

In this exercise you will learn to use several multiple alignment programs and compare the robustness of each. The exercise is on the Rieske iron sulfur protein from a number of species. The critical functional site of this protein is a iron-sulfur center with a bound [2Fe-2S] cluster. The amino acid binding motifs are known to have consensi of C-X-H-X-G-C and C-X-X-H. Evaluate the following alignment programs for the ability to discover the conserved motifs as well as to correctly align the rest of the protein sequences. The result you obtain may aid in the understanding of the origin and evolution of the respiratory process.

1. Retrieve the following protein sequences in the FASTA format using NCBI Entrez: P08067, P20788, AAD55565, P08980, P23136, AAC84018, AAF02198.
2. Save all the sequences in a single file using a text editor such as *nedit*.
3. First, use a progressive alignment program Clustal to align the sequences. Submit the multisequence file to the ClustalW server (www.ch.embnet.org/software/ClustalW.html) for alignment using the default settings. Save the result in ClustalW format in a text file.
4. To evaluate quality of the alignment, visually inspect whether the key residues that form the iron–sulfur centers are correctly aligned and whether short gaps are scattered throughout the alignment. A more objective evaluation is to use a scoring approach. Go to a web server for alignment quality evaluation (http://igs-server.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi?action=Evaluate%20a%20Multiple%20Alignment::Regularstage1=1). Bookmark this site for later visits. Paste the alignment in Clustal format in the query box. Click on "Submit."
5. To view the result in the next page, click on the "score_html" link. The overall quality score is given in the top portion of the file. And the alignment quality is indicated by a color scheme. Record the quality score.
6. Align the same sequences using a profile-based algorithm MultAlin (http://prodes.toulouse.inra.fr/multalin/multalin.html) using the default parameters. Click the button "Start MultAlin." Save the results in the FASTA format first and convert it to the Clustal format for quality comparison.
7. Select the hyperlink for "Results as a FASTA file." Copy the alignment and open the link for Readseq (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi/). Paste

the FASTA alignment in the query box. Select the output sequence format as Clustal. Click "Submit."

8. Copy and paste the Clustal alignment in the quality evaluation server and record the score.
9. Submit the same unaligned sequences to a semi-exhaustive alignment program DCA (http://bibiserv.techfak.uni-bielefeld.de/dca/submission.html). Click on the "Submission" link on the right (in the green area) and paste the sequences in the query box. Select the output format as "FASTA." Click "Submit." Save the results for format conversion using Readseq. Do quality evaluation as above.
10. Do alignment with the same unaligned sequences using an iterative tree-based alignment program PRRN (http://prrn.ims.u-tokyo.ac.jp/). Select the output format as "FASTA." Select the "Copy and Paste" option and enter your e-mail address before submitting the alignment. Compare the quality of the alignment with other methods.
11. Finally, align the sequences using the T-Coffee server (www.ch.embnet.org/software/TCoffee.html). Score of alignment is directly presented in the HTML format. Record the score for comparison purposes.
12. Carefully compare the results from different methods. Can you identify the most reasonable alignment? Which method appears to be the best?

Hidden Markov Model Construction and Searches

This exercise is about building a hidden Markov model (HMM) profile and using it to search against a protein database.

1. Obtain the above sequence alignment from T-Coffee in the Clustal format.
2. Copy and paste the alignment file to the query box of the HMMbuild program for building an HMM profile (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_hmmbuild.html). Click "Submit."
3. You may receive an error message "Your clustalw alignment doesn't start with the sentence : CLUSTAL W (...) multiple sequence alignment". Replace the header (beginning line) of the input file with "CLUSTAL W (...) multiple sequence alignment". Click "Submit."
4. When the HMM profile is constructed, click on the link "PROFILE" to examine the result.
5. Choose HMMSEARCH and UniProt-SwissProt database before clicking "Submit." This process takes a few minutes. Once the search is complete, the database hits that match with the HMM are returned along with multiple alignment files of the database sequences.
6. You have options to build a new HMM profile or to extract the full database sequences. Click "HMMBUILD" at the bottom of the page. The HMM profile building can be iterated as many times as desired similar to PSI-BLAST. For the interest of time, we stop here.

Protein Motif Searches

1. Align four sequences of different lengths in a file named "zf.fasta" (downloadable from www.cambridge.org/us/catalogue/catalogue.asp?isbn=0521600820) using T-Coffee and DIALIGN2 (http://bibiserv.techfak.uni-bielefeld.de/dialign/submission.html).
2. Which of the programs is able to identify a zinc finger motif [C(X4)C(X12)H(X3)H]?
3. Verify the result with the INTERPRO motif search server (www.ebi.ac.uk/interpro/) by cutting and pasting each of the unaligned sequences, one at a time, to the query box. Submit the query and inspect the search result.
4. Retrieve the protein sequence AAD42764 from Entrez. Do motif search of this sequence using a number of search programs listed. Pay attention to statistical scores such as *E*-values, if available, as well as the boundaries of the domains/motifs.
   a) BLOCKS Impala Searcher (http://blocks.fhcrc.org/blocks/impala.html).
   b) Reverse PSI-BLAST (http://blocks.fhcrc.org/blocks-bin/rpsblast.html).
   c) ProDom (http://prodes.toulouse.inra.fr/prodom/current/html/form.php).
   d) SMART (http://smart.embl-heidelberg.de/), select the "Normal" mode and paste the sequence in the query window.
   e) InterPro (www.ebi.ac.uk/interpro/), choose the link "Sequence Search" in the left grey area. Paste the sequence in the following query page.
   f) Scansite (http://scansite.mit.edu/), in the Motif Scan section, select "Scan a Protein by Input Sequence." Enter a name for the sequence and paste the sequence in the query window. Click "Submit Request." In the following page with the graphic representation of the domains and motifs, click "DOMAIN INFO" to get a more detailed description.
   g) eMatrix (http://fold.stanford.edu/ematrix/ematrix-search.html).
   h) Elm (http://elm.eu.org/). Use *Homo sapiens* as default organism.

Compile the results. What is your overall conclusion of the presence of domains and motifs in this protein?

DNA Motif Searches

DNA motifs are normally very subtle and can only be detected using "alignment-independent" methods such as expectation maximization (EM) and Gibbs motif sampling approaches.

1. Use the DNA sequence file "sd.fasta" (downloadable from www.cambridge.org/us/catalogue/catalogue.asp?isbn=0521600820) and generate alignment using the EM-based program Improbizer (www.cse.ucsc.edu/~kent/improbizer/improbizer.html) with default parameters.
2. Do the same search using a Gibbs sampling-based algorithm AlignAce (http://atlas.med.harvard.edu/cgi-bin/alignace.pl) using default parameters.

3. Compare the results of best scored motifs from both methods. Are there overlaps?
4. Copy and paste the first motif derived from AlignAce to *nedit*. Remove the illegal characters (spaces and numbers).
5. Cut and paste the motif alignment into the WebLogo program (http://weblogo.berkeley.edu/logo.cgi). Click the "Create logo" button.
6. Can you identify the bacterial Shine-Dalgarno sequence motif from the sequences?

EXERCISE 4. PHYLOGENETIC ANALYSIS

In this exercise, you will reconstruct the phylogeny of HIV by building an unrooted tree for the HIV/SIV gp120 proteins using the distance neighbor joining, maximum parsimony, maximum likelihood, and Bayesian inference methods.

Constructing and Refining a Multiple Sequence Alignment

1. Open the file "gp120.fasta" (downloadable from www.cambridge.org/us/catalogue/catalogue.asp?isbn=0521600820) using nedit.
2. Go to the MultAlin alignment server (http://prodes.toulouse.inra.fr/multalin/multalin.html). Copy and paste the sequences to the query box and submit the sequences for alignment using the default parameters.
3. Visually inspect the alignment result and pay attention to the matching of cysteine residues, which roughly indicate the correctness of the alignment.
4. View the result in FASTA format by clicking the hyperlink "Results as a fasta file." Save the FASTA alignment in a new text file using nedit.
5. Refine the alignment using the Rascal program that realigns certain portion of the file. Open the Rascal web page (http://igbmc.u-strasbg.fr/PipeAlign/Rascal/rascal.html) and upload the previous alignment file in FASTA format.
6. After a minute or so, the realignment is displayed in the next window. Examine the new alignment. If you accept the refinement, save the alignment in FASTA format.
7. Next, use the Gblocks program to further eliminate poorly aligned positions and divergent regions to make the alignment more suitable for phylogenetic analysis. Go to the Gblocks web page (http://molevol.ibmb.csic.es/Gblocks_server/index.html) and upload the above refined alignment into the server.
8. By default, the program should be set for protein sequences. Check the three boxes that allow less stringent criteria for truncation. These three boxes are "Allow smaller final blocks," "Allow gap positions within the final blocks," and "Allow less strict flanking positions."
9. Click the "Get Blocks" button. After the program analyzes the alignment quality, conserved regions are indicated with blue bars.

10. If you accept the selection, click on the "Resulting alignment" hyperlink at the bottom of the page to get selected sequence blocks in the FASTA format.
11. Copy the sequence alignment and change its format using the Readseq program (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi) by pasting the sequences to the query box of the program. Select "Phylip|Phylip4" as output format. Click submit. Save the final alignment in a scratch file.

Constructing a Distance-Based Phylogenetic Tree

1. Go to the WebPhylip web page (http://biocore.unl.edu/WEBPHYLIP/).
2. Select "Distance Computation" in the left window. In the subsequent window, select "PAM Matrix" under "Protein Sequences."
3. Copy and paste the above Phylip alignment in the query box on the lower right portion of the window. Leave everything else as default. Click the "Submit" button.
4. Once the distance matrix is computed, a split window on the upper right is refreshed to give the distance matrix of the dataset.
5. To construct a tree with the matrix, select "Run" the distance methods in the lower left window.
6. In the next window, select "Run" under "Neighbor-joining and UPGMA methods."
7. This refreshes the lower right window. By default, the "Neighbor-joining tree" is selected. Select "Yes" for the question "Use previous data set?" (highlighted in red). Click the "Submit" button.
8. A crude diagram of the phylogenetic tree is displayed in the upper right window.
9. To draw a better tree, select the "Draw trees" option (in green) in the left window.
10. In the next window, select "Run" under "Draw Cladograms and phenograms."
11. In the refreshed lower right window, make sure "Yes" is selected for the question "Use tree file from last stage?" Click the "Submit" button.
12. A postscript file is returned. Save it to hard drive as "filename.ps." Convert the postscript file to the PDF format using the command "ps2pdf filename.ps". Open the PDF file using the xpdf filename.pdf command.
13. Using the same alignment file, do a phylogenetic tree using the Fitch–Margoliash method and compare the final result with the neighbor-joining tree.

Constructing a Maximum Parsimony Tree

1. In the same WebPhylip web page, click "Back to menu."
2. Select "Protein" in the "Run phylogeny methods for" section in the left window.
3. Choose "Run" "Parsimony" in the next window.
4. In the refreshed window on the right, repaste the sequence alignment in the query window and choose "Yes" for "Randomize input order of sequences?"
5. Leave everything else unchanged and click the Submit button. This initiates the calculation for parsimony tree construction, which will take a few minutes.
6. Two equally most parsimonious trees are shown in the upper right window.

7. Choose "Do consensus" on the left and "Run" "Consensus tree" in the next window.
8. In the refreshed lower right window, make sure "Yes" is selected for the question "Use tree file from last stage?" Click the "Submit" button.
9. Choose "Draw trees" and then "Run" for "Draw Cladogram" on the left.
10. Make sure "Yes" for "Use tree file from last stage?" is selected and leave everything else as default.
11. A postscript image is returned for viewing.

Constructing a Quartet Puzzling Tree

1. Access the Puzzle program web page (http://bioweb.pasteur.fr/seqanal/interfaces/Puzzle.html).
2. Copy and paste (or upload) the gp120 Phylip alignment into the query window.
3. Select "protein" for the sequence type.
4. Scroll down the window to the Protein Options section. Select "JTT model" for amino acid substitutions.
5. Leave other parameters as default. Provide your e-mail address before submitting the query.
6. The URL for the results will be sent to you by e-mail (check your e-mail in about 10 or 20 minutes).
7. Get your result by following the URL in e-mail. Select the "drawgram" option and click "Run the selected program on results.tree" button.
8. In the following Phylip page, choose "Phenogram" as "Tree Style" in the next page, and click "Run drawgram."
9. The tree is returned in a postscript file. Open the image file by clicking on the hyperlink plotfile.ps.

Constructing a Maximum Likelihood Tree Using Genetic Algorithm

1. Go to the PHYML web page (http://atgc.lirmm.fr/phyml/).
2. Select "File" next to the query window. Click "Browse" to select the sequence alignment file for uploading.
3. Select "Amino-Acids" for "Data Type."
4. Leave everything else as default and provide your name, country, and e-mail address before submitting the query.
5. The treeing result will be sent to you by e-mail (the process takes about 10 or 20 minutes).
6. One of your e-mail attachment files should contain the final tree in the Newick format which can be displayed using the Drawtree program (www.phylodiversity.net/~rick/drawtree/).
7. Copy and paste the Newick file to the query window. Leave everything else as default. Click the "Draw Tree" button.
8. The graphical tree is returned in the PDF format.

Constructing a Phylogenetic Tree Using Bayesian Inference

1. Access the BAMBE program (http://bioweb.pasteur.fr/seqanal/interfaces/bambe.html).
2. Copy and paste (or upload) the gp120 alignment in Phylip format into the query window. Leave all parameters as default (6,000 cycles with 1,000 cycles as burn-in). Provide your e-mail address before submitting the query.
3. The URL of the result is returned via e-mail (in 10 or 20 minutes).
4. Verify that the likelihood value of the final tree has reached near convergence by checking the end of the "results.par" file. (In this file, the first column represents the number of cycles and second column the lnL values of the intermediate trees.)
5. If the log likelihood of the trees is indeed stabilized, go back to the previous page and draw the final consensus tree by selecting the "drawgram" option and clicking the "Run the selected program on results.tre" button.
6. In the following page, choose "Phenogram" as Tree Style in the next page, and click "Run drawgram."
7. The tree is returned in a postscript file. Open the image file by clicking on the hyperlink plotfile.ps.
8. Compare the phylogenetic results from different methods and draw a consensus of the results and consider the evolutionary implications. What is the phylogenetic relationship of HIV-1, HIV-2, and SIV? What do the trees tell you about the origin of HIV and how many events of cross-species transmissions? (*Note:* SIVCZ is from chimpanzee and SIVM is from macaque/mangabeys.)

EXERCISE 5. PROTEIN STRUCTURE PREDICTION

Protein Secondary Structure Prediction

In this exercise, use several web programs to predict the secondary structure of a globular protein and a membrane protein, both of which have known crystal structures. The predictions are used to compare with experimentally determined structures so you can get an idea of the accuracy of the prediction programs.

1. Retrieve the protein sequence YjeE (accession number ZP_00321401) in the FASTA format from NCBI Entrez. Download the sequence into a text file.
2. Predict its secondary structure using the GOR method. Go to the web page http://fasta.bioch.virginia.edu/fasta_www/garnier.htm and paste the sequence into the query box. Click the "Predict" button.
3. Save the result in a text file.
4. The crystal structure of this protein has a PDB code 1htw (as a homotrimeric complex). The secondary structure of each monomer can be retrieved at the PDBsum database (www.ebi.ac.uk/thornton-srv/databases/pdbsum/).
5. Enter the PDB code 1htw in the query box. Click "Find."

6. In the menu bar on the top right, select the "Protein" menu. This brings up the protein secondary structure as well as the CATH classification information. If the background of the secondary structure window is black, select and download the PDF file format next to the window. Open it using the xpdf command.
7. Compare the real secondary structure with the GOR prediction. What conclusion can you draw from this?
8. Now do another prediction using the neural-network–based Predator program (http://bioweb.pasteur.fr/seqanal/interfaces/predator-simple.html). Enter your e-mail address before submitting the query.
9. The result is returned in the "predator.out" file. Compare the result with the GOR prediction and the known secondary structure. What conclusion can you draw from this?
10. Do the structure prediction again using the BRNN-based Porter program (http://distill.ucd.ie/porter//).
11. Paste sequence in the query window and enter the e-mail address. Click the "Predict" button.
12. The result is e-mailed to you in a few minutes.
13. Compare the result with the previous predictions and the known secondary structure. What can you learn from this?
14. Retrieve the human aquaporin sequence (AAH22486) from NCBI.
15. Predict the transmembrane structure using the Phobius program (http://phobius.cgb.ki.se/). Record the result.
16. The PDB code of this protein structure is 1h6i, which you can use to retrieve the experimentally determined secondary structure from PDBsum.
17. Compare the prediction result with the known structure. Do the total number of transmembrane helices and their boundaries match in all cases?

Protein Homology Modeling

In the following exercise, construct a homology model for a small protein from a cyanobacterium, *Anabaena variabilis*. The protein, which is called HetY, may be involved in nitrogen fixation but has no well-defined function. The objective of this exercise is to help provide some functional clues of the protein. The protein model is displayed and rendered using a shareware program Chimera (downloadable from www.cgl.ucsf.edu/chimera/).

1. Retrieve the protein sequence (ZP_00161818) in the FASTA format from NCBI Entrez. Save the sequence in a text file.
2. To search for structure templates, do a BLASTP search (www.ncbi.nlm.nih.gov/BLAST/) against the "pdb" database.
3. Examine the BLAST result. Select the top hit if the *E*-value of the alignment is significant enough. This sequence should correspond to the structure 1htw and can serve can as the structure template.

4. Perform more refined alignment between HetY and the template. Click on the hyperlink in the header of the template sequence to retrieve the full-length sequence in the FASTA format. Save it in a text file.
5. Align HetY and the template sequence (1htw) using T-Coffee (www.ch.embnet.org/software/TCoffee.html).
6. Convert the alignment into the FASTA format using Readseq (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi/). Save it into a text file.
7. Refine the alignment using the Rascal server (http://igbmc.u-strasbg.fr/PipeAlign/Rascal/rascal.html) by uploading the FASTA alignment file.
8. Download the refined alignment in the FASTA format.
9. Perform comprehensive homology modeling using the GetAtoms server (www.softberry.com/berry.phtml?topic=getatoms&group=programs&subgroup=propt).
10. Paste the alignment in the query window. Select "FASTA" for format. Enter "1htw" for PDB identifier and "A" for chain identifier. Make sure the input order is the target sequence before the template sequence.
11. Select "Add H-atoms at the end of optimization" and "Process loops and insertions." Click the "PROCESS" button.
12. The coordinates of the model are returned in the next page. Save the coordinate file using the "Save As" option in File menu.
13. Open the coordinate file using nedit. Delete the dashes, trademark, and other HTML-related characters at the end of the file.
14. The raw model can be refined by energy minimization. Upload the edited coordinate file to the Hmod3DMM program (www.softberry.com/berry.phtml?topic=molmech&group=programs&subgroup=propt). Press "START."
15. A refined model is returned in a few minutes. Save the energy-minimized coordinates to the hard drive.
16. Check the quality of final structure using Verify3D (http://nihserver.mbi.ucla.edu/Verify_3D/). Upload the structure to the Verify3D server. Click "Send File."
17. In the resulting quality plot, scores above 0 indicate favorable conformations. Check to see whether if any residue scores are below 0. If the scores are significantly below 0, reminimization of the model is required.
18. Assuming the modeled protein is final, the next step is to add cofactor to the protein.
19. Assuming that the target protein has similar biochemical functions as the template protein (ATPase), important ligands from the template file that can be transferred to the target protein.
20. Download the template structure (1htw) from the PDB website (www.rcsb.org/pdb/). Click the "Download/Display File" link in the menu on the left. Download the noncompressed PDB file.
21. To extract the cofactor, open the 1HTW.pdb with nedit. Go the HETATM section near the bottom of the file. Find the coordinates for ADP 560. Copy the coordinates (make sure you include all the atoms for this cofactor).

22. Open the HetY model using nedit and paste the HETATM coordinates immediately after the ATOM section (near the end of the file). Delete the dashes, trademark, and other HTML-related characters.
23. Before using Chimera to visualize the model, you need add an alias to your .cshrc file. Open .cshrc with nedit and add a line at the end of the file: alias chimera /chem/ chimera/chimera-1.2065/bin/chimera. Save the file and quit nedit. In the UNIX window, type source .cshrc.
24. Invoke the Chimera program by typing chimera.
25. In the File menu, select "Open." For File type, select "all (ask type)." Select your model (e.g., hetY.pdb) to be opened.
26. The structure file is initially uncolored. Color the atoms by going to the menu Actions → Color → by element.
27. The structure can be rotated using the left mouse button, moved using the middle mouse button, and zoomed using right mouse button.
28. You can display a smooth solid surface showing electrostatic distribution as well as bound cofactor ADP. Go to Actions → Surface → show.
29. To select the cofactor, go to Select → Chain → het. To color it, select Actions → Color → cyan. To render it in spheres, select Actions → Atoms/bonds → sphere.
30. To finalize selection, go to Select → Clear selection. Rotate the model around to study the protein-ligand binding.
31. To reset the rendering, go to Actions → Surface → hide; Actions → Atoms/bonds → wire.
32. Now draw the secondary structure of the model. Select Actions → Ribbon → show; Actions → Ribbon → round. To color it, go to Tools → Graphics → Color Secondary Structure. In the pop-up window, click OK for default setting.
33. To hide the wire frames, select Actions → Atoms/bonds → hide.
34. To show the cofactor, Select → Chain → het. Then Actions → Atoms/bonds → ball & stick.
35. The publication quality image can be saved for printing purposes. To make it more printer-friendly, the background can be changed to white. Select Actions → Color → background; Actions → Color → white.
36. To save the image, go to File → Save Image. In the pop-up window, click "Save As."
37. As the program is writing an image, the model may dance around on the screen for a while. When it stabilizes, a new window pops up to prompt you for filename and file type (default format is .png). Give it a name and click "Save."
38. Quit the Chimera program. To view the image in a UNIX window, type imgview filename.png.
39. The image file can be e-mailed to yourself as attachment for printing on your own printer.
40. When you are done, close all the programs and log out.

EXERCISE 6. GENE AND PROMOTER PREDICTION AND GENE ANNOTATION

Gene Prediction

In this exercise, you do gene predictions using a bacterial sequence from *Heliobacillus mobilis* (Hm_dna.fasta) (downloadable from www.cambridge.org/us/catalogue/catalogue.asp?isbn=0521600820). This provides the foundation for operon predictions and promoter predictions. One way to verify the gene prediction result is to check the presence of Shine–Dalgarno sequence in front of each gene which is a purine-rich region with a consensus AGGAGG and is located within 20 bp upstream of the start codon.

1. Point your browser to the GeneMark web page (frame-by-frame module) (http://opal.biology.gatech.edu/GeneMark/fbf.cgi).
2. Upload the Hm_dna.fasta sequence file and choose *Bacillus subtilis* as "Species" (the closest organism).
3. Leave other options as default and start the GeneMark program.
4. Save the prediction result using nedit.
5. To confirm the prediction result, the sequence needs to have numbering.
6. Convert the original sequence file into a GenBank format using the ReadSeq server (http://iubio.bio.indiana.edu/cgi-bin/readseq.cgi). Save the results in a new file.
7. Based on the prediction by GeneMark, find the gene start sites in the sequence file. Can you find the Shine–Dalgarno sequence in each predicted frame?
8. Do another gene prediction using the Glimmer program (http://compbio.ornl.gov/GP3/pro.html). Select "Glimmer Genes." Use *B. subtilis* as the closest organism. Upload the sequence file and perform the Glimmer search.
9. When the data processing is complete, click the "Get Summary" button. In the following page, select Retrieve: → TextTable.
10. Compare the prediction result with that from GeneMark. Pay attention to the boundaries of open reading frames. For varied gene predictions, verify the presence of Shine–Dalgarno sequence in each case. Have you noticed problems of overpredictions or missed predictions with Glimmer? Can you explain why?

Operon Prediction

In this exercise, you predict operons of the above heliobacterial sequence using the 40-bp rule: if intergenic distance of a pair of unidirectionally transcribed genes is smaller than 40 bp, the gene pair can be called an *operon*. This rule was used widely before the development of the scoring method of Wang et al., which is a little too complicated for this lab.

1. Using the gene prediction result from GeneMark, calculate the intergenic distance of each pair of genes.
2. How many operons can you derive based on the 40-bp rule?

Promoter Prediction

In this exercise, perform ab initio promoter predictions based on the operon prediction from the previous exercise. Algorithms for promoter prediction are often written to predict the transcription start sites (TSS) instead. The −10 and −35 boxes can be subsequently deduced from the upstream region of this site.

1. Using the operon prediction result that you believe is correct, copy ∼150-bp upstream sequence from the first operon start sites and save the sequence in a new file.
2. Convert the sequence to the FASTA format, using the Readseq program.
3. Do a promoter prediction using BPROM (www.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb). Paste the sequence in the query window and press the PROCESS button. Record the result. (*Note:* TSS predicted by this program is labeled as LDF.)
4. Do another promoter prediction using the SAK program (http://nostradamus.cs.rhul.ac.uk/%7Eleo/sak_demo/), which calculates the likelihood scores of sites being the TSS.
5. In the output page, find the position that has the highest likelihood score (listed in the second column), which is the TSS prediction.
6. Compare the results from the two sets of predictions. Are they consistent?

Gene Annotation

A major issue in genomics is gene annotation. Although a large number of genes and proteins can be assigned functions simply by sequence similarity, about 40% to 50% of the genes from newly sequenced genomes have no known functions and can only be annotated as encoding "hypothetical proteins." In this exercise, you are given one of such "difficult" protein sequences for functional annotation. This protein is YciE from *E. coli,* which has been implicated in stress response. However, its actual biochemical function has remained elusive. In this exercise, use advanced bioinformatics tools to derive functional information of the protein sequence.

1. Retrieve the protein sequence of YciE from NCBI Entrez (www.ncbi.nlm.nih.gov/Entrez/, accession P21363) in the FASTA format and study the existing annotation in the GenBank file.
2. Do domain and motif searches of this sequence using RPS-BLAST (http://blocks.fhcrc.org/blocks/rpsblast.html), SMART (http://smart.embl-heidelberg.de/) (Use the Normal mode. Check all four boxes below the query box for Outlier homologs, PFAM domains, signal peptides, and internal repeats before starting the search) and InterPro (www.ebi.ac.uk/interpro/). Compile the results. What is the consensus domains and motifs in this protein?
3. Do a functional prediction based on protein interactions by using the STRING server (http://string.embl.de/). Paste your sequence into the query box and click the "GO" button.

4. In the result page for predicted protein–protein interactions, check to see what are the predicted interacting proteins. What is the evidence for the interaction prediction?
5. Do a protein threading analysis using the HHPred server (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred). This program searches protein folds by combining HMMs and secondary structure prediction information.
6. Paste the sequence in the query box and submit the job using all the default settings. The query is processed. The result is returned in a minute.
7. Pick the top hit showing the most significant *E*-value and study the annotation of the structure match and visually inspect the alignment result.
8. Get more detailed information about the best matching protein structure by clicking the link with the PDB code. This brings up the PDB beta page with detailed annotation information especially the bibliographic information of the structure.
9. You can retrieve the original publication on the structure by selecting the "PubMed" link. Read the "Introduction" section of this paper. Can you get any functional description about the protein in that paper in relation to the stress response?
10. In the PDB site, retrieve the sequence of this structure (only one subunit) by selecting the menu "Summarize" → "Sequence Details." In the next page, scroll down the window to click the "Download" button for chain A in FASTA format.
11. Open the sequence in the FASTA format and save it to the hard disk.
12. Do a refined pairwise alignment of this sequence with YciE using the AliBee server (www.genebee.msu.su/services/malign_reduced.html). What is the percent identity for the best set of alignment?
13. Do a PRSS test on the two original sequences by copying and pasting the two unaligned sequences in two individual query boxes of the PRSS server (http://fasta.bioch.virginia.edu/fasta/prss.htm). Select 1,000 shuffles with "window" setting. Are the two protein sequences significantly related? Can you designate them as homologous sequences?
14. Compile the results from each of the predictions. What is your overall conclusion of the function of this protein?

APPENDIX TWO

Glossary

Ab initio prediction: computational prediction based on first principles or using the most elementary information.

Accession number: unique number given to an entry in a biological database, which serves as a permanent identifier for the entry.

Agglomerative clustering: microarray data clustering method that begins by first clustering the two most similar data points and subsequently repeating the process to merge groups of data successively according to similarity until all groups of data are merged. This is in principle similar to the UPGMA phylogenetic approach.

Alternative splicing: mRNA splicing event that joins different exons from a single gene to form variable transcripts. This is one of the mechanisms of generating a large diversity of gene products in eukaryotes.

Bayesian analysis: statistical method using the Bayes theorem to describe conditional probabilities of an event. It makes inferences based on initial expectation and existing observations. Mathematically, it calculates the posterior probability (revised expectation) of two joint events (A and B) as the product of the prior probability of A event given the condition B (initial expectation) and conditional probability of B (observation) divided by the total probability of event A with and without the condition B. The method has wide applications in bioinformatics from sequence alignment and phylogenetic tree construction to microarray data analysis.

Bioinformatics: discipline of storing and analyzing biological data using computational techniques. More specifically, it is the analysis of the sequence, structure, and function of the biological macromolecules – DNA, RNA, and proteins – with the aid of computational tools that include computer hardware, software, and the Internet.

Bit score: statistical indicator in database sequence similarity searches. It is a normalized pairwise alignment score that is independent of database size. It is suitable for comparing search results from different databases. The higher the bit score, the better the match is.

BLAST (Basic Local Alignment Search Tool)**:** commonly used sequence database search program based on sequence similarity. It has many variants, such as BLASTN, BLASTP, and BLASTX, for dealing with different types of sequences. The major feature of the algorithm is its search speed, because it is designed to rapidly detect a region of local sequence similarity in a database sequence and use it as anchor to extend to a fuller pairwise alignment.

BLOSUM matrix: amino acid substitution matrix constructed from observed frequencies of substitution in blocks of ungapped alignment of closely related protein sequences. The numbering of the BLOSUM matrices corresponds to percent identity of the protein sequences in the blocks.

Boolean expression: database retrieval method of expressing a query by connecting query words using the logical operators AND, OR, and NOT between the words.

Bootstrap analysis: statistical method for assessing the consistency of phylogenetic tree topologies based on the generation of a large number of replicates with slight modifications in input data. The trees constructed from the datasets with random modifications give a distribution of tree topologies that allow statistical assessment of each individual clade on the trees.

CASP (Critical Assessment in Structure Prediction)**:** biannual international contest to assess protein structure prediction software programs using blind testing. This experiment attempts to serve as a rigorous test bed by providing contestants with newly solved but unpublished proteins structures to test the efficacy of new prediction algorithms. By avoiding the use of known protein structures as benchmarks, the contest is able to provide unbiased assessment of the performance of prediction programs.

Chromosome walking: experimental technique that identifies overlapping genomic DNA clones by labeling the ends of the clones with oligonucleotide probes. Through a multistep process, it is able cover an entire chromosome.

Clade: group of taxa on a phylogenetic tree that are descended from a single common ancestor. They are also referred to as being *monophyletic*.

COG (Cluster of Orthologous Groups)**:** protein family database based on phylogenetic classification. It is constructed by comparing protein sequences encoded by completely sequenced genomes and identifying orthologous proteins shared by three or more genomes to be clustered together as orthologous groups.

Comparative genomics: subarea of genomics that focuses on comparison of whole genomes from different organisms. It includes comparison of gene number, gene location, and gene content from these genomes. The comparison provides insight into the mechanism of genome evolution and gene transfer among genomes.

Contig: contiguous stretch of DNA sequence assembled from individual overlapping DNA segments.

Cytological maps: maps showing banding patterns on a stained chromosome and observed under a microscope. The bands are often associated with the locations of genetic markers. The distance between any two bands is expressed in relative units (Dustin units).

Database: computerized archive used for storage and organization of data in such a way that information can be retrieved easily via a variety of search criteria.

Divisive clustering: microarray data clustering method that works by lumping all data points in a single cluster and successively dividing the data into smaller groups according to similarity until all the hierarchical levels are resolved. This is in principle similar to the neighbor joining phylogenetic approach.

DNA microarray: technology for high throughput gene expression profiling. Oligonucleotides representing every gene in a genome can be immobilized on tiny spots on the surface of a glass chip, which can be used for hybridization with a labeled cDNA population. By analyzing the hybridization result, levels of gene expression at the whole genome level can be revealed.

Domain: evolutionarily conserved sequence region that corresponds to a structurally independent three-dimensional unit associated with a particular functional role. It is usually much larger than a motif.

Dot plot: visual technique to perform a pairwise sequence alignment by using a two-dimensional matrix with each sequence on its dimensions and applying dots for matching residues. A contiguous line of dots in a diagonal indicates a local alignment.

Dynamic programming: algorithm to find an optimal solution by decomposing a problem into many smaller, sequentially dependent subproblems and solving them individually while storing the intermediate solutions in a table so that the highest scored solution can be chosen. To perform a pairwise sequence alignment, the method builds a two-dimensional matrix with each sequence on its dimensions and applies a scoring scheme to fill the matrix and finds the maximum scored region representing the best alignment by backtracking through the matrix.

EM (expectation maximization)**:** local multiple sequence alignment method for identification of shared motifs among input sequences. The motifs are discovered through random alignment of the sequences to produce a trial PSSM and successively refinement of the PSSM. A motif can be recruited after this process is repeated many times until there is no further improvement on the matrix.

EST (expressed sequence tags)**:** short sequences obtained from cDNA clones serving as short identifiers of full length genes. ESTs are typically in the range of 200 to 400 nucleotides in length and are generated using a high throughput approach. EST profiling can be used as a snapshot of gene expression in a particular tissue at a particular stage.

***E*-value** (expectation value)**:** statistical significance measure of database sequence matches. It indicates the probability of a database match expected as a result of random chance. The *E*-value depends on the database size. The lower the *E*-value, the more significant the match is.

Exon shuffling: mRNA splicing event that joins exons from different genes to generate more transcripts. This is one of the mechanisms of generating a large diversity of gene products in eukaryotes.

FASTA: database sequence search program that performs the pairwise alignment by employing a heuristic method. It works by rapidly scanning a sequence to identify identical words of a certain size as the query sequence and subsequently searching for regions that contain a high density of words with high scores. The high-scoring regions are subsequently linked to form a longer gapped alignment, which is later refined using dynamic programming.

False negative: true match that fails to be recognized by an algorithm.

False positive: false match that is incorrectly identified as a true match by an algorithm.

Fingerprint: group of short, ungapped sequence segments associated with diagnostic features of a protein family. A fingerprint is a smaller unit than a motif.

Flat file: database file format that is a long text file containing database entries separated by a delimiter, a special character such as a vertical bar (|). Each field within an entry is separated by tabs.

Fold: three-dimensional topology of a protein structure described by the arrangement and connection of secondary structure elements in three dimensional space.

Fold recognition: method of protein structure prediction for the most likely protein structural fold based on structure profile similarity and sequence profile similarity. The structure profiles incorporate information of secondary structures and solvation energies. The term has been used interchangeably with *threading*.

Functional genomics: study of gene functions at the whole-genome level using high throughput approaches. This study is also termed *transcriptome analysis*, which refers to the analysis of the full set of RNA molecules produced by a cell under a given condition.

Gap penalty: part of a sequence alignment scoring system in which a penalizing score is used for producing gaps in alignment to account for the relative rarity of insertions and deletions in sequence evolution.

Gene annotation: process to identify gene locations in a newly sequenced genome and to assign functions to identified genes and gene products.

Gene ontology: annotation system for gene products using a set of structured, controlled vocabulary to indicate the biological process, molecular function, and cellular localization of a particular gene product.

Genetic algorithm: computational optimization strategy that performs iterative and randomized selection to achieve an optimal solution. It uses biological terminology as metaphor because it involves iterative "crossing," which is a mix and match of mathematical routines to generate new "offspring" routines. The offsprings are allowed to randomly "mutate." A scoring system is applied to select an offspring among many with a higher score (better "fitness") than the "parents." This offspring is allowed to "propagate" further. The iterations continue until the "fittest" offspring or an optimal solution is selected.

Genetic map: map of relative positions of genes in a genome, based on the frequency of recombinations of genetic markers through genetic crossing. The distance between two genetic markers is measured in relative units (Morgans).

Genome: complete DNA sequence of an organism that includes all the genetic information.

Genomics: study of genomes characterized by simultaneous analysis of all the genes in a genome. The topics of genomics range from genome mapping, sequencing, and functional genomic analysis to comparative genomic analysis.

Gibbs sampling: local multiple sequence alignment method for identification of shared motifs among input sequences. PSSMs are constructed iteratively from N–1 sequences and are refined with the left-out sequence. An optimal motif can be recruited after this process is repeated many times until there is no further improvement on the matrix.

Global alignment: sequence alignment strategy that matches up two or more sequences over their entire lengths. It is suitable for aligning sequences that are of similar length and suspected to have full-length similarity. If used for more divergent sequences, this strategy may miss local similar regions.

Heuristics: computational strategy to find a near-optimal solution by using rules of thumb. Essentially, this strategy takes shortcuts by reducing the search space according to certain criteria. The results are not guaranteed to be optimal, but this method is often used to save computational time.

Hidden Markov model: statistical model composed of a number of interconnected Markov chains with the capability to generate the probability value of an event by taking into account the influence from hidden variables. Mathematically, it calculates probability values of connected states among the Markov chains to find an optimal path within the network of states. It requires training to obtain the probability values of state transitions. When using a hidden Markov model to represent a multiple sequence alignment, a sequence can be generated through the model by incorporating probability values of match, insertion, and deletion states.

Hierarchical clustering: technique to classify genes from a gene expression profile. The classification is based on a gene distance matrix and groups genes of similar expression patterns to produce a dendrogram.

Hierarchical sequencing: sequencing approach that divides the genomic DNA into large fragments, each of which is cloned into a bacterial artificial chromosome (BAC). The relative order of the BAC clones are first mapped onto a chromosome. Each of the overlapping BAC clones is subsequently sequenced using the shotgun approach before they are assembled to form a contiguous genomic sequence.

Homologs: biological features that are similar owing to the fact that they are derived from a common ancestry.

Homology: biological similarity that is attributed to a common evolutionary origin.

Homology modeling: method for predicting the three-dimensional structure of a protein based on homology by assigning the structure of an unknown protein using an existing homologous protein structure as a template.

Homoplasy: observed sequence similarity that is a result of convergence or parallel evolution, but not direct evolution. This effect, which includes multiple substitutions at individual positions, often obscures the estimation of the true evolutionary distances between sequences and has to be corrected before phylogenetic tree construction.

HSP (high scoring segment pair)**:** intermediate gapless pairwise alignment in BLAST database sequence alignment.

Identity: quantitative measure of the proportion of exact matches in a pairwise or multiple sequence alignment.

Jackknife: tree evaluation method to assess the consistency of phylogenetic tree topologies by constructing new trees using only half of the sites in an original dataset. The method is similar to bootstrapping, but its advantages are that sites are not duplicated relative to the original dataset and that computing time is much reduced because of shorter sequences.

Jukes–Cantor model: substitution model for correcting multiple substitutions in molecular sequences. For DNA sequences, the model assumes that all nucleotides are substituted with an equal rate. It is also called the *one-parameter model*.

k-Means clustering: classification technique that identifies the association of genes in an expression profile. The classification first assigns data points randomly among a number of predefined clusters and then moves the data points among the clusters while calculating the distances of the data points to the center of the cluster (centroid). The process is iterated many times until a best fit of all data points within the clusters is reached.

Kimura model: substitution model for correcting multiple substitutions in molecular sequences. For DNA sequences, the model assumes that there are two different substitution rates, one for transition and the other for transversion. It is also called the *two-parameter model*.

Lateral gene transfer: process of gene acquisition through exchange between species in a way that is incongruent with the commonly accepted vertical evolutionary scenario. It is also called *horizontal gene transfer*.

Linear discriminant analysis: statistical method that separates true signals from background noise by projecting data points in a two-dimensional graph and drawing a diagonal line that best separates signals from nonsignals based on the patterns learned from training datasets.

Local alignment: pairwise sequence alignment strategy that emphasizes matching the most similar segments between the two sequences. It can be used for aligning sequences of significant divergence and unequal lengths.

Log-odds score: score that is derived from the logarithmic conversion of an observed frequency value of an event divided by the frequency expected by random chance so that the score represents the relative likelihood of the event. For example, a positive log-odds score indicates an event happens more likely than by random chance.

Low-complexity region: sequence region that contains a high proportion of redundant residues resulting in a biased composition that significantly differs from the general sequence composition. This region often leads to spurious matches in sequence alignment and has to be masked before being used in alignment or database searching.

Machine learning: computational approach to detect patterns by progressive optimization of the internal parameters of an algorithm.

Markov process: linear chain of individual events linked together by probability values so that the occurrence of one event (or state) depends on the occurrence of the previous event(s) (or states). It can be applied to biological sequences in which each character in a sequence can be considered a state in a Markov process.

Maximum likelihood: statistical method of choosing hypotheses based on the highest likelihood values. It is most useful in molecular phylogenetic tree construction.

Maximum parsimony: principle of choosing a solution with fewest explanations or logic steps. In phylogenetic analysis, the maximum parsimony method infers a tree with the fewest mutational steps.

Minimum evolution: phylogenetic tree construction method that chooses a tree with minimum overall branch lengths. In principle, it is similar to maximum parsimony, but differs in that the minimum evolution method is distance based, whereas maximum parsimony is character based.

Molecular clock: assumption that molecular sequences evolve at a constant rate. This implies that the evolutionary time of a lineage can be estimated from its branch length in a phylogenetic tree.

Molecular phylogenetics: study of evolutionary processes and phylogenies using DNA and protein sequence data.

Monophyletic: refers to taxa on a phylogenetic tree that are descended from a single common ancestor.

Monte Carlo procedure: computer algorithm that produces random numbers based on a particular statistical distribution.

Motif: short, conserved sequence associated with a distinct function.

Needleman–Wunsch algorithm: a global pairwise alignment algorithm that applies dynamic programming in a sequence alignment.

Negative selection: evolutionary process that does not favor amino acid replacement in a protein sequence. This happens when a protein function has been stabilized. The implied function constraint deems mutations to be deleterious to the

protein function. This can be detected when the synonymous substitution rate is higher than the nonsynonymous substitution rate in a protein encoding region.

Neighbor joining: phylogenetic tree-building method that constructs a tree based on phylogenetic distances between taxa. It first corrects unequal evolutionary rates of raw distances and uses the corrected distances to build a matrix. Tree construction begins from a completely unresolved tree and then decomposes the tree in a stepwise fashion until all taxa are resolved.

Neural network: machine-learning algorithm for pattern recognition. It is composed of input, hidden, and output layers. Units of information in each layer are called nodes. The nodes of different layers are interconnected to form a network analogous to a biological nervous system. Between the nodes are mathematical weight parameters that can be trained with known patterns so they can be used for later predictions. After training, the network is able to recognize correlation between an input and output.

Newick format: text representation of tree topology that uses a set of nested parentheses in which each internal node is represented by a pair of parentheses that enclose all members of a monophyletic group separated by a comma. If a tree is scaled, branch lengths are placed immediately after the name of the taxon separated by a colon.

Nonsynonymous substitutions: nucleotide changes in a protein coding region that results in alterations in the encoded amino acid sequences.

Object-oriented database: database that stores data as units that combine data and references to other records. The units are referred to as *objects*. Searching a such database involves navigating through the objects via pointers and links. The database structure is a more flexible than that of relational database but lacks the rigorous mathematical foundation of the relational databases.

OMIM (Online Mendelian Inheritance in Man)**:** database of human genetic disease, containing textual descriptions of the disorders and information about the genes associated with genetic disorders.

Orthologs: homologous sequences from different organisms or genomes derived from speciation events rather than gene duplication events.

Outgroup: taxon or a group of taxa in a phylogenetic tree known to have diverged earlier than the rest of the taxa in the tree and used to determine the position of the root.

Overfitting: phenomenon by which a machine learning algorithm overrepresents certain patterns while ignoring other possibilities. This phenomenon is a result of insufficient amounts of data in training the algorithm.

PAM matrix: amino acid substitution matrix describing the probability of one amino acid being substituted by another. It is constructed by first calculating the number of observed substitutions in a sequence dataset with 1% amino acid mutations and subsequently extrapolating the number of substitutions to more divergent sequence datasets through matrix duplication. The PAM unit is theoretically related to

evolutionary time, with one PAM unit corresponding to 10 million years of evolutionary changes. Thus the higher the PAM numbering, the more divergent amino acid sequences it reflects.

Paralogs: homologous sequences from the same organism or genome, which are derived from gene duplication events rather than speciation events.

Phylogenetic footprinting: process of finding conserved DNA elements through aligning DNA sequences from multiple related species. It is widely used for identifying regulatory elements in a genome.

Phylogenetic profile: the pattern of coexistence or co-absence of gene pairs across divergent genomes. The information is useful for making inference of functionally linked genes or genes encoding interacting proteins.

Phylogeny: study of evolutionary relationships between organisms by using treelike diagrams as representations.

Physical map: map of locations of gene markers constructed by using a chromosome walking technique. The distance between gene markers is measured directly as kilobases (Kb).

Positive selection: evolutionary process that favors the replacement of amino acids in a protein sequence. This happens when the protein is adapting to a new functional role. The evidence for positive selection often comes from the observation that the nonsynonymous substitution rate is higher than the synonymous substitution rate in the DNA coding region.

Posterior probability: probability of an event estimated after taking into account a new observation. It is used in Bayesian analysis.

Profile: scoring matrix that represents a multiple sequence alignment. It contains probability or frequency values of residues for each aligned position in the alignment including gaps. A weighting scheme is often applied to correct the probability for unobserved and underobserved sequence characters. Profiles can be used to search sequence databases to detect distant homologs. This term is often used interchangeably with *position-specific scoring matrix* (PSSM).

Progressive alignment: multiple sequence alignment strategy that uses a stepwise approach to assemble an alignment. It first performs all possible pairwise alignments using the dynamic programming approach and determines the relative distances between each pair of sequences to construct a distance matrix, which is subsequently used to build a guide tree. It then realigns the two most closely related sequences using the dynamic programming approach. Other sequences are progressively added to the alignment according to the degree of similarity suggested by the guide tree. The process proceeds until all sequences are used in building a multiple alignment. The Clustal program is a good example of applying this strategy.

Protein family: group of homologous proteins with a common structure and function. A protein family is normally constructed from protein sequences with an overall identity of at least 35%.

Proteome: complete set of proteins expressed in a cell.

Proteomics: study of a proteome, which involves simultaneous analyses of all translated proteins in the entire proteome. Its topics include large-scale identification and quantification of expressed proteins and determination of their localization, modifications, interactions, and functions.

PSI-BLAST: unique version of the BLAST program that employs an iterative database searching strategy to construct multiple sequence alignments and convert them to profiles that are used to detect distant sequence homologs.

PSSM (position-specific scoring matrix)**:** scoring table that lists the probability or frequency values of residues derived from each position in an ungapped multiple sequence alignment. A PSSM can be weighted or unweighted. In a weighted PSSM, a weighting scheme is applied to correct the probability for unobserved and underobserved sequence characters. This term is often used interchangeably with *profile*.

***P*-value:** statistical measure representing the significance of an event based on a chance distribution. It is calculated as the probability of an event supporting the null hypothesis. The smaller the *P*-value, the more unlikely an event is due to random chance (null hypothesis) and therefore the more statistically significant it is.

Quadratic discriminant analysis: statistical method that separates true signals from background noise by projecting data points in a two dimensional graph and by drawing a curved line that best separates signals from nonsignals based on knowledge learned from a training dataset.

Quartet puzzling: phylogenetic tree construction method that relies on compiling tree topologies of all possible groups of four taxa (quartets). Individual four-taxon trees are normally derived using the exhaustive maximum likelihood method. A final tree that includes all taxa is produced by deriving a consensus from all quartet trees. The advantage of this method is computational speed.

Query: specific value used to retrieve a particular record from a database.

Ramachandran plot: two-dimensional scatter plot showing torsion angles of each amino acid residue in a protein structure. The plot delineates preferred or allowed regions of the angles as well as disallowed regions based on known protein structures. This plot helps in the evaluation of the quality of a new protein model.

Relational database: database that uses a set of separate tables to organize database entries. Each table, also called *relation,* is made up of columns and rows. Columns represent individual fields and rows represent records of data. One or more columns in a table are indexed so they can be cross-referenced in other tables. To answer a query to a relational database, the system selects linked data items from different tables and combines the information into one report.

RMSD (root mean square deviation)**:** measure of similarity between protein structures. It is the square root of the sum of the squared deviations of the spatial coordinates of the corresponding atoms of two protein structures that have been superimposed.

Regular expression: representation format for a sequence motif, which includes positional information for conserved and partly conserved residues.

Rotamer: preferred side chain torsion angles based on the knowledge of known protein crystal structures.

Rotamer library: collection of preferred side chain conformations that contains information about the frequency of certain conformations. Having a rotamer library reduces the computational time in a side chain conformational search.

SAGE (serial analysis of gene expression)**:** high throughput approach to measure global gene expression patterns. It determines the quantities of transcripts by using a large number of unique short cDNA sequence tags to represent each gene in a genome. Compared to EST analysis, SAGE analysis has a better chance of detecting weakly expressed genes.

Scaffold: continuous stretch of DNA sequence that results from merging overlapping contigs during genome assembly. Scaffolds are unidirectionally oriented along a physical map of a chromosome.

Self-organizing map: classification technique that identifies the association of genes in an expression profile. The classification is based on a neural network–like algorithm that first projects data points in a two dimensional space and subsequently carries out iterative matching of data points with a predefined number of nodes, during which the distances of the data points to the center of the cluster (centroid) are calculated. The data points stay in a particular node if the distances are small enough. The iteration continues until all data points find a best fit within the nodes.

Sensitivity: measure of ability of a classification algorithm to distinguish true positives from all possible true features. It is quantified as the ratio of true positives to the sum of true positives plus false negatives.

Sequence logo: graphical representation of a multiple sequence alignment that displays a consensus sequence with frequency information. It contains stacked letters representing the occurrence of the residues in a particular column of a multiple alignment. The overall height of a logo position reflects how conserved the position is; the height of each letter in a position reflects the relative frequency of the residue in the alignment.

Shotgun sequencing: genome sequencing approach that breaks down genomic DNA into small clones and sequences them in a random fashion. The genome sequence is subsequently assembled by joining the random fragments after identifying overlaps.

Shuffle test: statistical test for pairwise sequence alignment carried out by allowing the order of characters in one of the two sequences to be randomly altered. The

shuffled sequence is subsequently used to align with the reference sequence using dynamic programming. A large number of such shuffled alignments serve to create a background alignment score distribution which is used to assess the statistical significance of the score of the original optimal pairwise alignment. A *P*-value is given to indicate the probability that the original alignment is a result of random chance.

Similarity: quantitative measure of the proportion of identical matches and conserved substitutions in a pairwise or multiple alignment.

Site: column of residues in a multiple sequence alignment.

Smith–Waterman algorithm: local pairwise alignment algorithm that applies dynamic programming in alignment.

Specificity: measure of ability of a classification algorithm to distinguish true positives from all predicted features. It is quantified as the ratio of true positives to the sum of true positives plus false positives.

Substitution matrix: two-dimensional matrix with score values describing the probability of one amino acid or nucleotide being replaced by another during sequence evolution. Commonly used substitution matrices are BLOSSUM and PAM.

Supervised classification: data analysis method that classifies data into a predefined set of categories.

Support vector machine: data classification method that projects data in a three-dimensional space. A "hyperplane" (a linear or nonlinear mathematical function) is used to separate true signals from noise. The algorithm requires training to be able to correctly recognize patterns of true features.

Synonymous substitutions: nucleotide changes in a protein coding sequence that do not result in amino acid sequence changes for the encoded protein because of redundancy in the genetic code.

Synteny: conserved gene order pattern across different genomes.

Systems biology: field of study that uses integrative approaches to model pathways and networks at the cellular level.

Taxon: each species or sequence represented at the tip of each branch of a phylogenetic tree. It is also called an *operational taxonomic unit* (OTU).

Threading: method of predicting the most likely protein structural fold based on secondary structure similarity with database structures and assessment of energies of the potential fold. The term has been used interchangeably with *fold recognition*.

Transcriptome: complete set of mRNA molecules produced by a cell under a given condition.

Transition: substitution of a purine by another purine or a pyrimidine by another pyrimidine.

Transversion: substitution of a purine by a pyrimidine or a pyrimidine by a purine.

True negative: false match that is correctly ignored by an algorithm.

True positive: true match that is correctly identified by an algorithm.

Unsupervised classification: data analysis method that does not assume predefined categories, but identifies data categories according to actual similarity patterns. It is also called *clustering*.

UPGMA (unweighted pair-group method with arithmetic means)**:** phylogenetic tree-building method that involves clustering taxa based on phylogenetic distances. The method assumes the taxa to have equal distance from the root and starts tree building by clustering the two most closely related taxa. This produces a reduced matrix, which allows the next nearest taxa to be added. Other taxa are sequentially added using the same principle.

***Z*-score:** statistical measure of the distance of a value from the mean of a score distribution, measured as the number of standard deviations.

Index

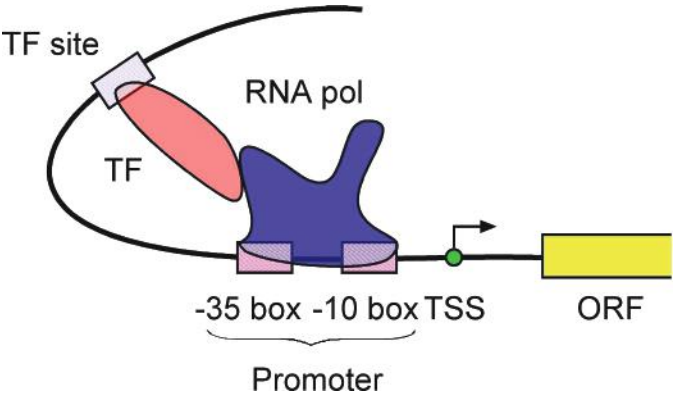


Figure 9.1. Schematic representation of elements involved in bacterial transcription initiation. RNA polymerase binds to the promoter region, which initiates transcription through interaction with transcription factors binding at different sites. *Abbreviations:* TSS, transcription start site; ORF, reading frame; pol, polymerase; TF, transcription factor (see page 114).

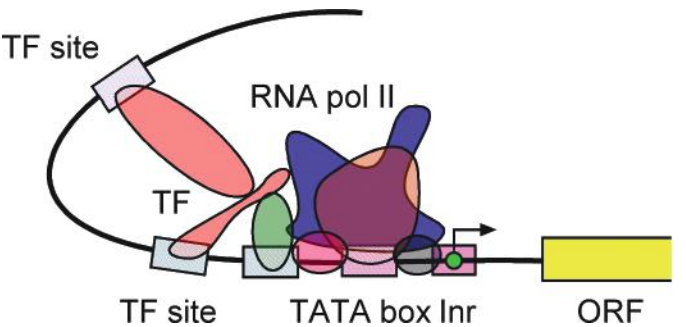


Figure 9.2. Schematic diagram of an eukaryotic promoter with transcription factors and RNA polymerase bound to the promoter. *Abbreviations:* Inr, initiator sequence; ORF, reading frame; pol, polymerase; TF, transcription factor (see page 115).

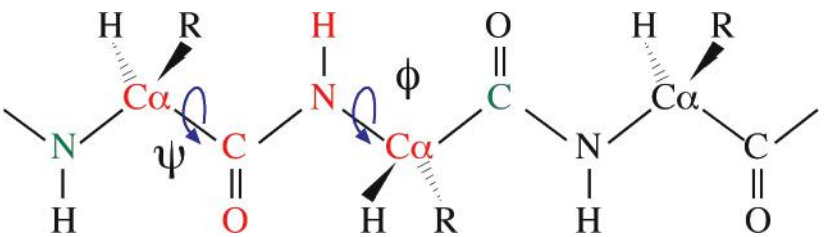


Figure 12.3. Definition of dihedral angles of ϕ and ψ. Six atoms around a peptide bond forming a peptide plane are colored in red. The ϕ angle is the rotation about the N–Cα bond, which is measured by the angle between a virtual plane formed by the C–N–Cα and the virtual plane by N–Cα–C (C in green). The ψ angle is the rotation about the Cα–C bond, which is measured by the angle between a virtual plane formed by the N–Cα–C (N in green) and the virtual plane by Cα–C–N (N in red) (see page 176).

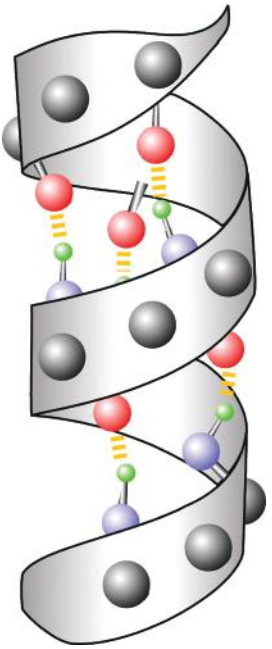


Figure 12.5. A ribbon diagram of an α-helix with main chain atoms (as gray balls) shown. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of two residues are shown in yellow dashed lines (see page 178).

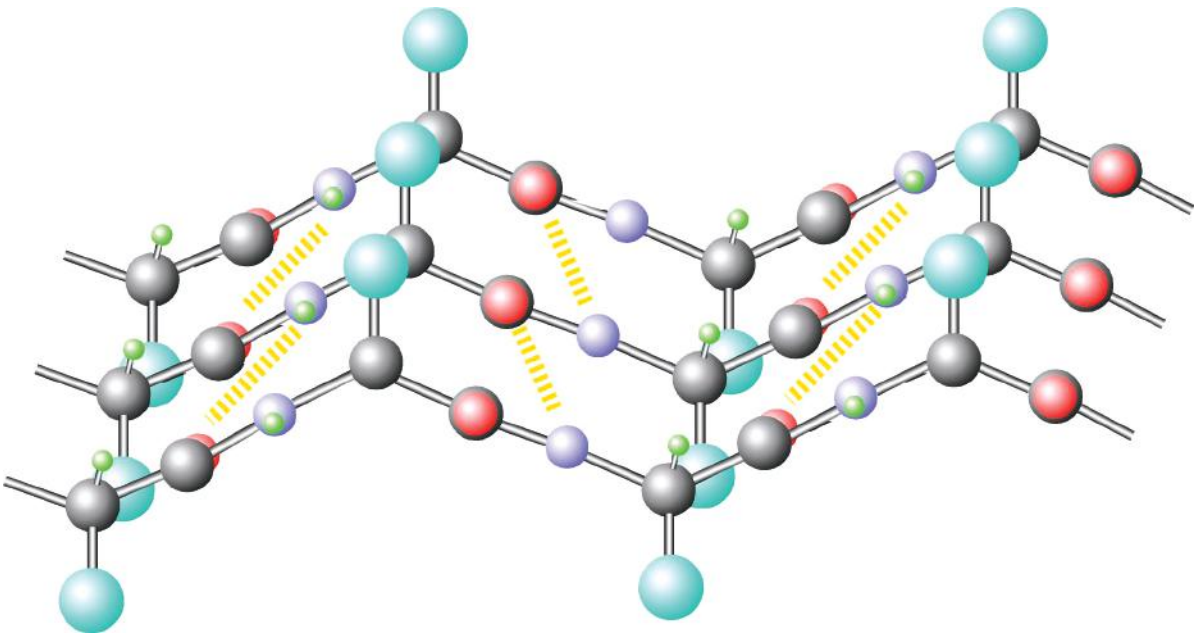


Figure 12.6. Side view of a parallel β-sheet. Hydrogen bonds between the carbonyl oxygen (red) and the amino hydrogen (green) of adjacent β-strands are shown in yellow dashed lines. R groups are shown as big balls in cyan and are positioned alternately on opposite sides of β-strands (see page 179).

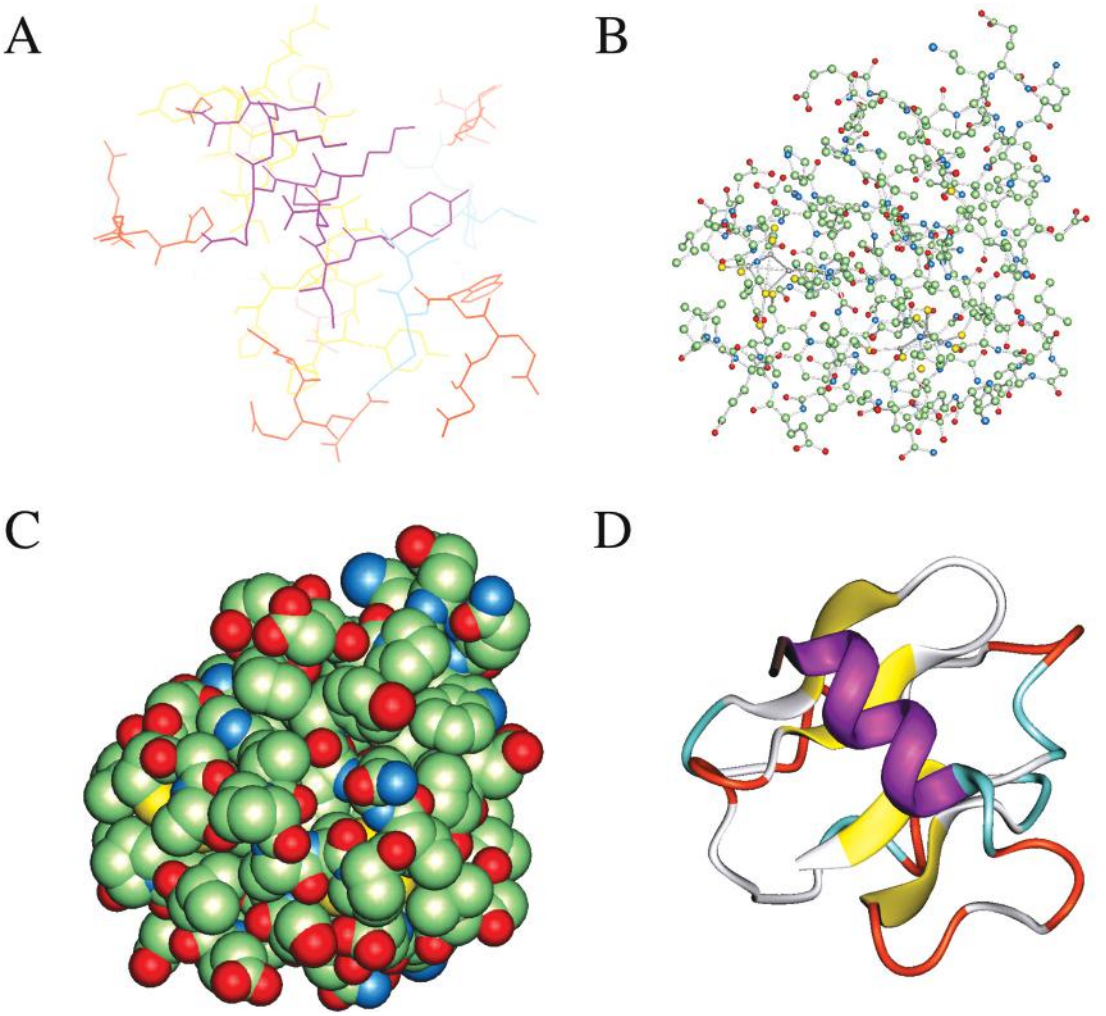


Figure 13.1. Examples of molecular structure visualization forms. **(A)** Wireframes. **(B)** Balls and sticks. **(C)** Space-filling spheres. **(D)** Ribbons (see page 188).

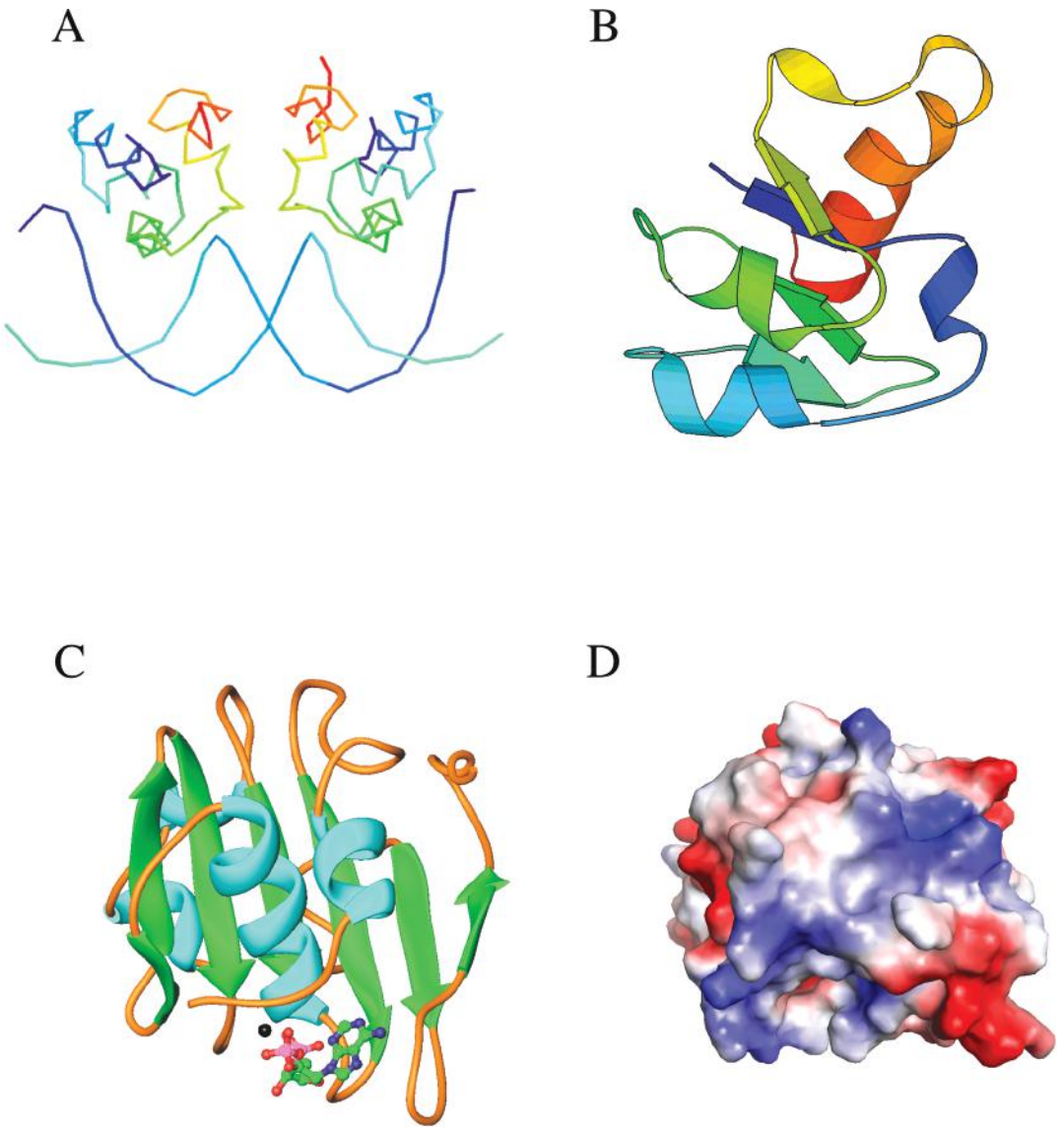


Figure 13.2. Examples of molecular graphic generated by **(A)** Rasmol, **(B)** Molscript, **(C)** Ribbons, and **(D)** Grasp (see page 189).

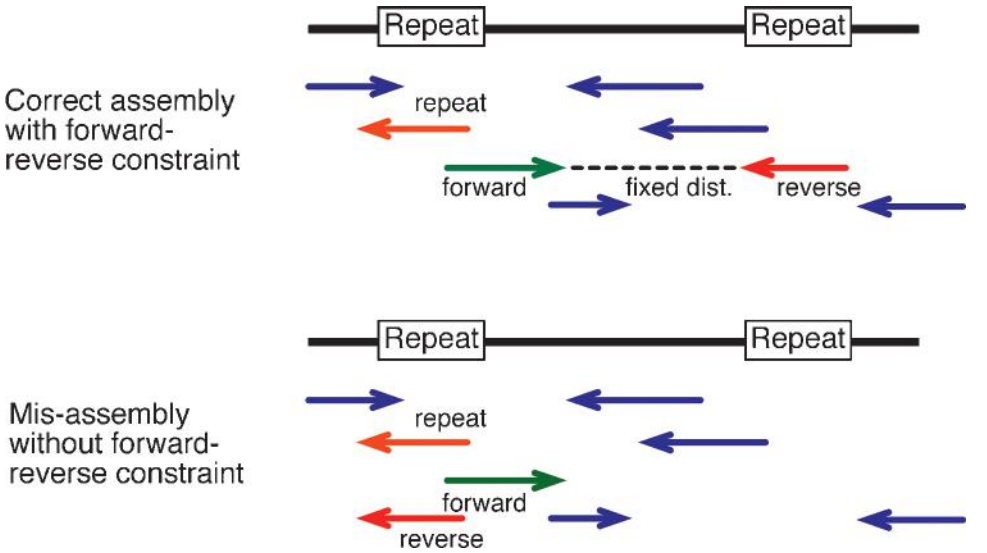


Figure 17.4. Example of sequence assembly with or without applying forward–reverse constraint, which fixes the sequence distance from both ends of a subclone. Without the restraint, the red fragment is misassembled due to matches of repetitive element in the middle of a fragment (see page 248).

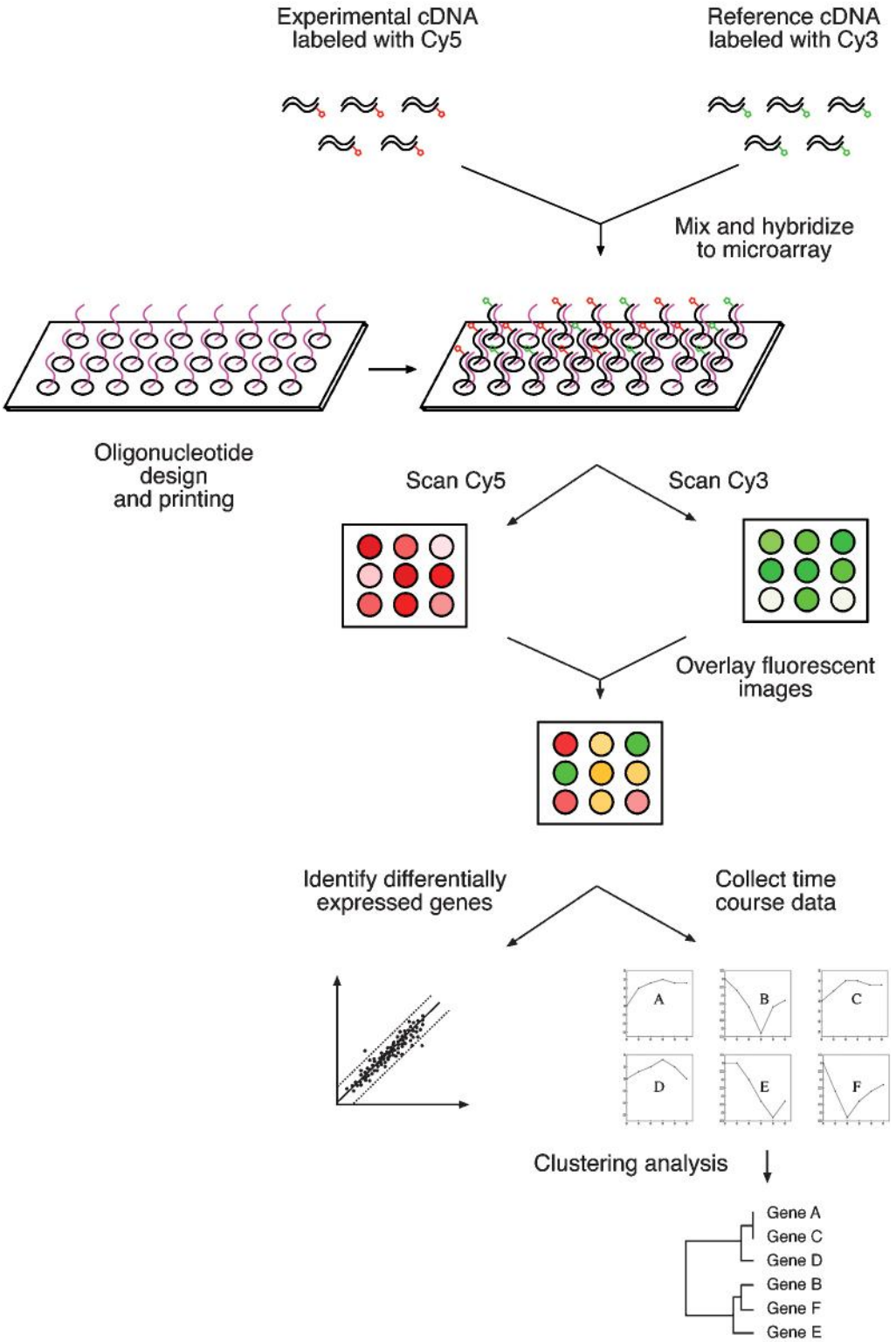


Figure 18.4. Schematic of a multistep procedure of a DNA microarray assay experiment and subsequent data analysis (see page 268).

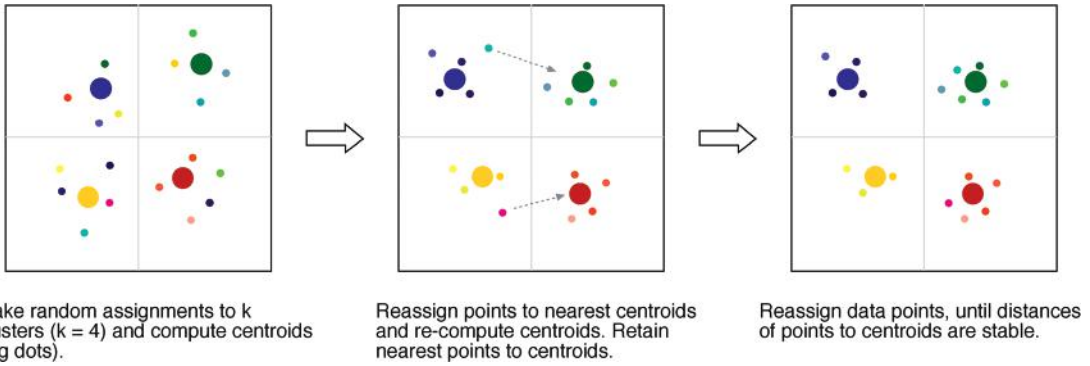
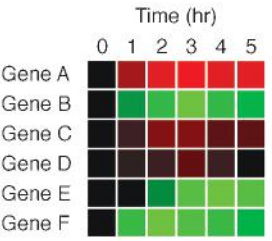


Figure 18.7. Example of k-means clustering using four partitions. Closeness of data points is indicated by resemblance of colors (see page 277).

	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	1	4	6	8	6	6
Gene B	1	0.6	0.3	0.1	0.3	0.4
Gene C	1	2	4	4	3	3
Gene D	1	1.5	2	3	2	1
Gene E	1	1	0.5	0.2	0.1	0.2
Gene F	1	0.3	0.1	0.2	0.3	0.4

convert to false colors

$\log_2$ conversion



	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	-0.82	0.96	0.65	-0.68	-0.79
Gene B		-0.85	-0.86	0.66	0.67
Gene C			0.70	-0.65	-0.87
Gene D				-0.41	-0.72
Gene E					0.26

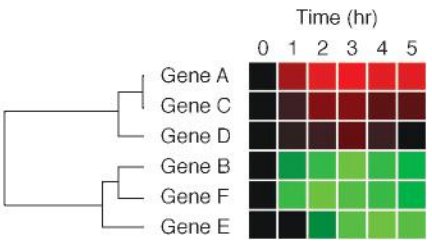
calculating Pearson correlation coefficients between genes

	0 hr	1 hr	2 hr	3 hr	4 hr	5 hr
Gene A	0	2	2.6	3	2.6	2.6
Gene B	0	-0.7	-1.7	-3.3	-1.7	-1.3
Gene C	0	1	2	2	1.6	1.6
Gene D	0	0.6	1	1.6	1	0
Gene E	0	0	-1	-2.3	-3.3	-2.3
Gene F	0	-1.7	-3.3	-2.3	-1.7	-1.3

conversion of coefficients to positive distance values

	Gene B	Gene C	Gene D	Gene E	Gene F
Gene A	1.82	0.04	0.35	1.68	1.79
Gene B		1.85	1.86	0.34	0.33
Gene C			0.30	1.65	1.87
Gene D				1.41	1.72
Gene E					0.74

hierarchical clustering



Box 18.1. Outline of the Procedure for Microarray Data Ananlysis (see page 271).

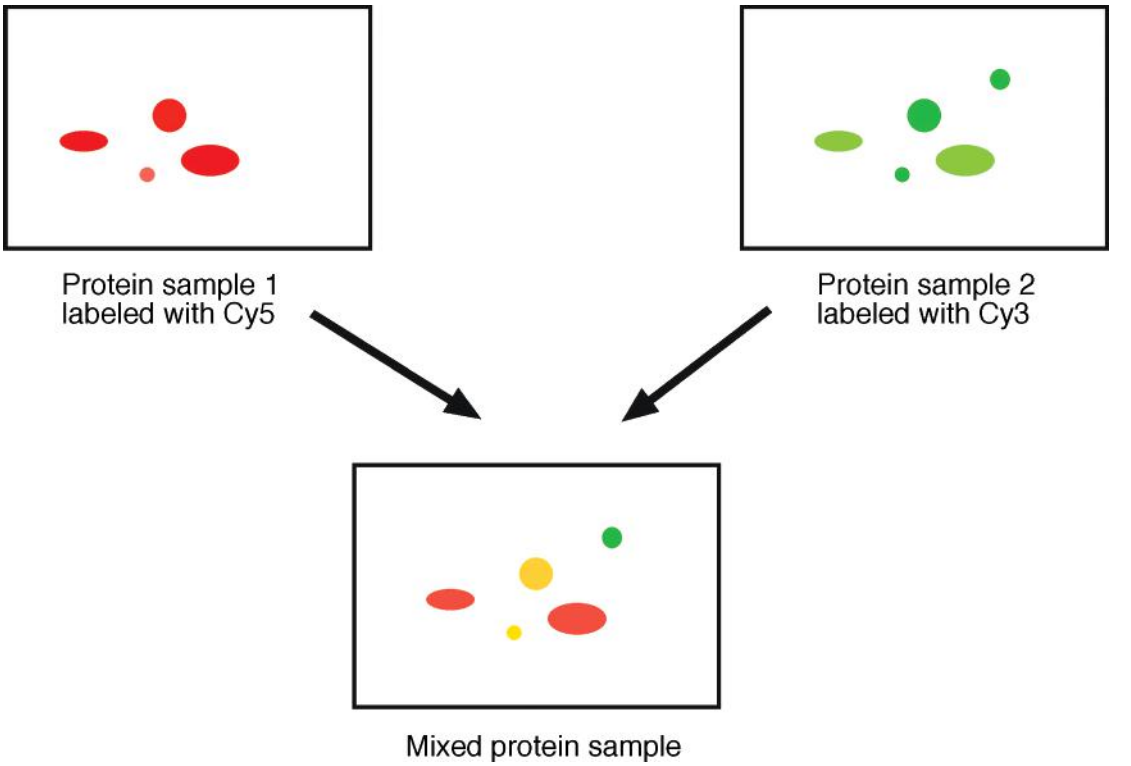


Figure 19.2. Schematic diagram showing protein differential detection using DIGE. Protein sample 1 (representing experimental condition) is labeled with a red fluorescent dye (Cy5). Protein sample 2 (representing control condition) is labeled with a green fluorescent dye (Cy3). The two samples are mixed together before running on a two-dimensional gel to obtain a total protein differential display map (see page 286).