BIOINFORMATICS
Unit V

Proteomics is the a branch of biology concerned with applying the techniques of molecular biology, biochemistry, and genetics to analyzing the structure, function, and interactions of the proteins produced by the genes of a particular cell, tissue, or organism, with organizing the information in databases, and with applications of the data.

Proteomics is the analysis of the entire protein complement of a cell, tissue, or organism under a specific, defined set of conditions. In its present state, it is dependent on decades of technological and instrumental developments.

Proteins are directly involved in almost every biological process, so comprehensive analysis of the proteins in the cell provides a unique global perspective on how these molecules interact and cooperate to create and maintain a working biological system. The cell responds to internal and external changes by regulating the level and activity of its proteins, so changes in the proteome, either qualitative or quantitative, provide a snapshot of this regulatory network in action.

The proteome is a complex and dynamic entity that can be defined in terms of the sequence, structure, abundance, localization, modification, interaction, and biochemical function of each of its components, providing a rich and varied source of data. The study of the proteome raises a number of potential ethical issues, such as those concerning the ownership, storage, and use of human tissues; the storage and use of data arising from proteomic research (especially if this affects donor privacy or could lead to discrimination); the extent to which informed consent is required; and questions regarding intellectual property and the use of human samples for proteomic research that later results in a commercial product. The analysis of the diverse properties of the proteome requires an equally diverse range of technologies as well as methods for data integration and mining, which further clouds the issue of ownership and intellectual property. Proteomics provides a much more robust and representative picture of the functioning cell than do other forms of large-scale biology, such as genome sequencing or the global analysis of gene expression; therefore, the potential ethical risks associated with sample and data misuse are greater.

Computational Resources for Protein Structure prediction

One of the key challenges in protein science is determining three dimensional structure from amino acid sequence. Although experimental methods for determining protein structures are providing high resolution structures, they cannot keep the pace at which amino acid sequences are resolved on the scale of entire genomes. Various computational tools have been developed that predict different levels of protein structural hierarchy. List of important tools for the prediction of protein structure is given below.

Protein structure prediction (more correctly called Protein inference) is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).

The most successful techniques for prediction of the protein three dimensional structures rely on aligning the sequence of a protein of unknown structure to a homolog of known structure. Such methods fail if there is no homolog in the structural database, or if the technique for searching the structural database is unable to identify homologs that are present. While absence of a homolog must await further X-ray or NMR structures, up to 4/5 of known homologues may be missed even by the best conventional pairwise sequence comparison methods.

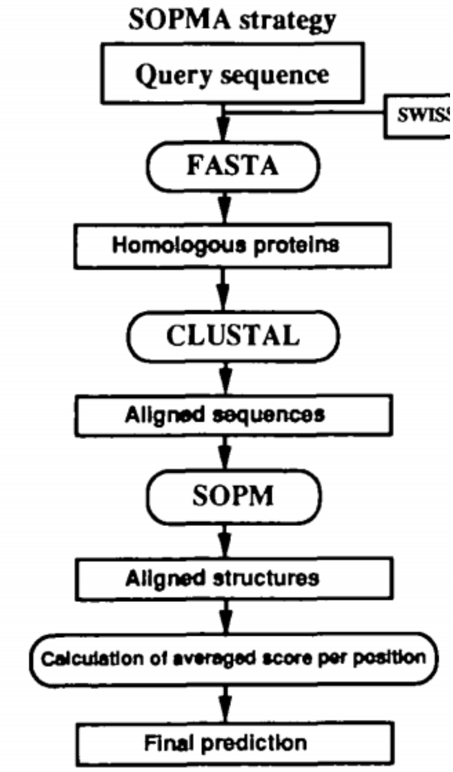
Techniques that exploit evolutionary information from protein families3-e or use empirical pair-potentialsl0'11 can normally detect more homologs than pairwise sequence comparison methods. An even greater challenge is to detect proteins that share similar folds, but are not clearly derived from a common ancestor (e.g. Rossman fold domains of lactate dehydrogenase and glycogen phosphorylase, and SH2-BirA12). Techniques for the prediction of protein secondary structure provide information that is useful both in ab initio structure prediction and as an additional constraint for fold-recognition algorithms.

However, for all these applications it is essential that the secondary structure prediction be accurate, or at least that, the reliability for each residue can be assessed. The majority of secondary structure prediction algorithms derive parameters or rules from an analysis of proteins of known three dimensional structure. The parameters are then applied by the algorithm to the sequence of unknown structure. Such approaches rely on having sufficient data to obtain reliable parameters and to avoid over-training for a specific data set.

Protein structure has four levels that define several aspects of protein structure: the primary structure, the secondary structure, tertiary structure, and quaternary structure. The primary structure of a protein is simply the sequence of amino acids that is translated from a messenger RNA(mRNA). The secondary structure of the protein is sub-regions of the primary structure that begin to interact and form alpha-helices(α-helices) and beta-sheets(β-sheets). The tertiary structure results when α-helices and βsheets within a single protein to form a three-dimensional shape. The final level of protein structure is quaternary structure, which refers to when more than one protein come together to form a complex. An example of a quaternary structure is human hemoglobin, which is made up of four distinct subunits, each an individual chain of amino acids, but functions as a single complex. A protein's final structure is deteremined by its inherent properties and when it becomes stable in a low energy state. In some cases, a chaperone helps another protein fold by introducing a lower energy barrier and shapes the protein into a conformation that the protein would not take on itself under physiological conditions.

SOPMA

Self-optimized prediction method is based on the homologue method. This method correctly predicts 69.5% of **amino acid**s for a three description of the secondary structure (α-helix, β-sheet and random coil) in a whole database containing 126 chains of non-homologous proteins on combination of SOPMA and PHD methods correctly predicts 82.2% of residues.



JPred

JPred is a Protein Secondary Structure Prediction server and has been in operation since approximately 1998. JPred incorporates the Jnet algorithm in order to make more accurate predictions. In addition to **protein secondary structure JPred also makes predictions on Solvent Accessibility and Coiled-coil regions (Lupas method).**

The current version of JPred (v4) has the following improvements and updates incorporated:

- Retrained on the latest UniRef90 and SCOPe/ASTRAL version of Jnet (v2.3.1) - mean secondary structure prediction accuracy of >82%
- Upgraded the Web Server to the latest technologies (Bootstrap framework, JavaScript) and updating the web pages – improving the design and usability through implementing responsive technologies.
- Upgraded the results reporting – both, on the web-site, and through the optional email summary reports: improved batch submission, added results summary preview through Jalview results visualization summary in SVG and adding full multiple sequence alignments into the reports.
- Improved help-pages, incorporating tool-tips, and adding one-page step-by-step tutorials.

Protein Structure visualization using RASMOL

RasMol is an important scientific tool for visualisation of molecules created by Roger Sayle in 1992. RasMol is used by hundreds of thousands of users world-wide to view macromolecules and to prepare publication-quality images.

RasMol Features

RasMol is a molecular graphics program intended for the visualisation of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. RasMol runs on wide range of architectures and operating systems including Microsoft Windows, Apple Macintosh, UNIX and VMS systems.

The program reads in a molecule coordinate file and interactively displays the molecule on the screen in a variety of colour schemes and molecule representations. Currently available representations include depth-cued wireframes, 'Dreiding' sticks, spacefilling (CPK) spheres, ball and stick, solid and strand biomolecular ribbons, atom labels and dot surfaces.

The program reads in molecular coordinate files and interactively displays the molecule on the screen in a variety of representations and colour schemes. Supported input file formats include Protein Data Bank (PDB), Tripos Associates' Alchemy and Sybyl Mol2 formats, Molecular Design Limited's (MDL) Mol file format, Minnesota Supercomputer Center's (MSC) XYZ (XMol) format, CHARMm format, CIF format and mmCIF format files. If connectivity information is not contained in the file this is calculated automatically.

The loaded molecule can be shown as wireframe bonds, cylinder 'Dreiding' stick bonds, alpha-carbon trace, space-filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands), hydrogen bonding and dot surface representations. Atoms may also be labelled with arbitrary text strings. Alternate conformers and multiple NMR models may be specially coloured and identified in atom labels.

Different parts of the molecule may be represented and coloured independently of the rest of the molecule or displayed in several representations simultaneously. The displayed molecule may be rotated, translated, zoomed and z-clipped (slabbed) interactively using either the mouse, the scroll bars, the command line or an attached dial box. RasMol can read a prepared list of commands from a 'script' file (or via inter-process communication) to allow a given image or viewpoint to be restored quickly. RasMol can also create a script file containing the commands required to regenerate the current image. Finally, the rendered image may be written out in a variety of formats including either raster or vector PostScript, GIF, PPM, BMP, PICT, Sun rasterfile or as a MolScript input script or Kinemage.

The RasMol help facility can be accessed by typing "help <topic>" or "help <topic> <subtopic>" from the command line. A complete list of RasMol commands may be displayed by typing "help commands". A single question mark may also be used to abbreviate the keyword "help". Please type "help notices" for important notices.

Running RasMol Under Microsoft Windows

To start RasMol under Microsoft Windows, double click on the RasMol icon in the program manager. When RasMol first starts, the program displays a single main window (the display window) with a black background on the screen and provides the command line window minimized as a small icon at the bottom of the screen. The command line or terminal window may be opened by double clicking on this RasMol icon.

It is possible to specify either a coordinate filename or a script filename or both on the windows command line. A script file may be specified by adding the option '-script <filename>' to the command line. A molecule coordinate file may be specified by placing its name on the command line, optionally preceded by a file format option. If no format option is given, the specified coordinate file is assumed to be in PDB, CIF or mmCIF format. If both a coordinate file and a script file are specified on the command line, the molecule is loaded first, then the script commands are applied to it. If either file is not found, the program displays the error message 'Error: File not found!' and the user is presented the RasMol prompt.

It is also possible to specify the initial graphics window size or position or both the size and the position with the options '-height nnnn', '-width nnnn', '-xpos nnnn' and '-ypos nnnn'. The numeric values are in pixels. The position is specified in terms of the top left corner of the rendering area.

RasMol's Window

On all platforms RasMol displays two windows, the main **graphics or canvas window** with a black background and **a command line or terminal window**. At the top of the graphics window (or at the top of the screen for the

Macintosh) is the RasMol menu bar. The contents of the menu bar change from platform to platform to support the local user interface guidelines; however, all platforms support the 'File', 'Display', 'Colours', 'Export', 'Options' and 'Settings' pull-down menus. The Main graphics window also has two scroll bars, one on the right and one at the bottom, that may be used to rotate the molecule interactively.

While the mouse pointer is located within the graphics area of the main display window, the mouse pointer is drawn as a cross-hair cursor, to enable the 'picking' of objects being displayed; otherwise the mouse pointer is drawn as an arrowhead. Any characters that are typed at the keyboard while the display window is in 'focus' (meaning active or foreground) are redirected to the command line in the terminal window. Hence you do not need continually to switch focus between the command line and graphics windows.

The display window may be resized at any point during the session. This has the effect of simply rescaling the image displayed on the canvas. RasMol imposes limits on the size of the display window such that the window must be large enough to display the menu and scroll bars and yet small enough to fit on a single screen. Attempts to enlarge the screen may fail owing to insufficient memory on the host machine, in which case RasMol reports the error message 'Renderer Error: Unable to allocate frame buffer!' or some similar error.

Mouse Controls

Here is a summary of RasMol's mouse click-and-drag controls. The 'set mouse' command mode defaults to 'set mouse rasmol', which gives the controls summarized below.

Action	Windows
Rotate X, Y	Left
Translate X, Y	Right
Rotate Z	Shift-Right
Zoom	Shift-Left
Slab Plane	Ctrl-Left

Command Line Interface

RasMol allows the execution of interactive commands typed at the RasMol prompt in the terminal window. Characters typed into either the terminal or the display window are processed on the command line. Each command must be given on a separate line terminated by a newline or carriage return character.. If a command is not recognised by RasMol, the program will generate an 'Unrecognised command!' error and redisplay the main prompt. If surplus information is given at the end of a command line, RasMol will execute the recognised command, but issue the warning message 'Warning: Ignoring rest of command!'. Some commands may prompt the user for more information. These commands display a different prompt and are discussed in the command reference.

Using RasMol and the PDB to Visualize Protein Structure

Program

RasMol is a molecular graphics program intended for the visualization of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images.

Rasmol is installed on your own computer if you have one.

Objective

After completing this laboratory, you should know how to:

Search the protein databank for protein structures

Download the primary structure (text format)

Manipulate proteins using molecular visualization software

Identify amino acids in an active site

Protein Visualization

The first step is to find the US hosted version of the PDB

Opening a PDB file and identifying the primary structure of a protein

1. One of the proteins we will be investigating is cytochrome c.

Using advanced search select keyword and search for "oxidized C2 cytochrome" in the databank (don't include quote marks).

2. Find PDB identifier 2C2C

Download the file for cytochrome c (oxidized form)

There are 4 icons under the PDB identifier

Click on the icon that looks like text

This is a preview of the data file

The information should look like this: (click here) It is of value to do this to make sure it is the file that you want to down load.

3. Now download the entire PDB file for viewing.
Click on the Icon with an arrow. The file can Either be saved to disk or opened directly
*You may also choose to look at the molecule using a PDB viewer (such as CHIME) by selecting [PDB viewer] or RasMol by selecting [Motifs-RasMol].
You will open a saved PDB file manually in RasMol.

Viewing the Molecule in RasMol

5. Open RasMol by clicking on the icon Raswin.exe (or RasMol).
If you prefer to launch RasMol from the Molecules R US website select:
Output requested [*Motifs-RasMol]*
For the first time use of RasMol on a computer, Netscape (or Explorer) must know where to find the application. Select browse and find the folder containing *Raswin.exe* and click on it. The molecule should now be opened in RasMol. Future use will not require these steps.
RasMol contains two parts: a viewing window and a command line.
Viewing window: displays the molecule
Command line: enables the user to change the view by typing commands

6. Open the molecule by clicking **File: Open** and selecting the PDB file saved in your network account.
A wireframe molecule will appear in the viewing window.
You will need to view the molecule and command line for this exercise.
Expand the viewing window so that it fills the opper ¾ of the screen.
Expand the command window so that it fills the remaining ¼ of the screen.

Sometimes when Rasmol opens the command line is minimized on the lower task bar. If you can't find the command line window look along the bottom of your screen and click on the Rasmol type there.

7. Manipulate the molecule using the mouse:
Translation - Right mouse button
Rotation - Left mouse button
Rotation in a plane - Shift - Right mouse button
Zoom Shift - left mouse button

8. Locate residues (amino acids) and particular atoms by using the mouse to pick or select atoms (i.e., click on an atom and the result is shown in the command window). The command window tells you which atoms are being selected and to what amino acid they belong. If you can't see the amino acid selected zoom in using the mouse. If you don't have the primary sequence in front of you type **show sequence** in the command line to see the sequence in the command line.

Note: many of the following tasks can be accomplished by typing the correct command into the command line or accessing the task descriptor from the pull down menus. You may do either to accomplish the task. Many commands, however, are only available through the command line. The command line allows greater flexibility and detail control.

9. Change the viewing type by selecting **Display** from the viewing window menu bar.
If only certain residues are changing, then type **select all (or use pull down menu).**
Practice changing display types for the protein this can be accomplished using the pull down menu.

10. Display particular amino acids:
Type **select all** and choose **Display** - *wireframe*
Type **select ala** to look at all of the alanines in the protein (75 atoms should be selected)
Type **color red** (all alalines should be red)
To view alaline as ball and stick select **Display** - *Ball and Stick.*
You can select any amino acid (again, type show sequence if you need to know what amino acids are present in the molecule). Practice looking for the loctions of various amino acids by selecting them one at a time and changing their color and Display style.
Typical colors: Red, orange, yellow, green, blue, violet, purple, brown, gold, cyan, black, white, grey, greentint, greenblue,hotpink, magenta, pink, pinktint, redorange, skyblue, bluetint, yellowtint.
Other colors are generated by specifying the **R**ed**G**reen**B**lue triplet values as in **color[255,0,0]** for red; **color [0,255,0]** for green; and **color [0,0,255]** for blue and all values in between.
Change the molecule's color to purple by typing **select all** and then **color [255,0,255].** When you are finished, type **select all,** and reset the **display** to *wireframe* and color cpk**.**

11. Display particular atoms:
Type **select all** and then **wireframe** (this is an alternative way Of changing the display)
Cytochrome c contains an iron metal center - Type select iron to identify the iron in the protein.
Type **color orange** (iron should be orange).
View it as a ball and stick model.
The iron is in a porphyrin ring. Identify the porphyrin ring by typing **select ligand.**
Change its color to red and display it as a ball and stick .

Type **select iron** and change its color to *yellow.*
Type **select backbone** to select the protein backbone.
Change the display to ribbon by typing **ribbon.**
Change the backbone color to grey.
Select the sidechains by typing **select sidechain.**
Change the color to *greentint.*
Change the background color to white by typing **background white.**
Practice changing styles and colors and rotate and move the

Copy and paste your final picture into Word or Wordperfect to be included in your report. If possible print in color; if no indicate what colors were present on the screen in pencil on your printout.

12. You can select individual amino acids by typing **select #** (i.e., **select 12** will select amino acid number 12; **select 1-12** will select residues 1-12).
This is useful for identifying domains (structurally independent units in a protein) or to highlight regions of interest in a protein sequence (i.e., beta sheets, helices, turns).

HOMOLOGY MODELING USING SWISS-MODEL

Homology modeling aims to build three-dimensional protein structure models using experimentally determined structures of related family members as templates. SWISS-MODEL workspace is an integrated Web-based modeling expert system. For a given target protein, a library of experimental protein structures is searched to identify suitable templates. On the basis of a sequence alignment between the target protein and the template structure, a three-dimensional model for the target protein is generated. Model quality assessment tools are used to estimate the reliability of the resulting models. Homology modeling is currently the most accurate computational method to generate reliable structural models and is routinely used in many biological applications. Typically, the computational effort for a modeling project is less than 2 h. However, this does not include the time required for visualization and interpretation of the model, which may vary depending on personal experience working with protein structures.

Homology modeling with swiss-model

The SWISS-MODEL is a simple and popular homology-modelling program and one of only few which available on the Internet. It uses the “building by fragments” method to construct the model on the template structures.
1. Open Swiss-Model at http://swissmodel.expasy.org/.

2. Link to **First Approach mode** (the upper link on the left frame).

3. The first data that we should supply (apart of the personal details) is the primary sequence of the protein we wish to model. In this exercise we will model the structure of a cyclic AMP dependent kinase (PKA). In order to get the primary sequence of this kinase you can enter to the ***Swissprot*** site (http://www.expasy.ch/sprot/), to type the accession number (P05132) and retrieve the entry. Save the sequence into local file in Fasta format. This new file should be later opened as a Fasta format file, meaning that it should begin with description line starting with the character ">". Call this file as **pka.tfa**. Copy and paste this sequence also to the relevant window at the SWISS-MODEL form (without the description line). This is actually all you need to do in order to run SWISS-MODEL in simple first approach mode. However, we **will not run** the program in this mode, instead we will run in **First approach mode with a specific template**. We will supply a specific kinase structures that will serve as templates during the building. These will be the structures of two tyrosine protein kinases (PDB ID 1iep, chain A and 1k2p, chain A). Under "**Use a specific template**" insert: 1iepA. Send the request.

4. We will now learn how to run the program in optimized mode. **Open Swiss-PDB-viewer.**

5. Choose "**Load Raw Sequence to Model**" item of the "Swiss Model" menu to load the file **pka.tfa** that you previously created.

6. Choose the "Swiss-Model" item of the "Preferences" menu. Enter your name and e-mail address. Make sure that the address of the modeling server is: **http://swissmodel.expasy.org/cgi-bin/sm-submit-request.cgi** and that the address of the template server is: **http://swissmodel.expasy.org/cgi-bin/blastexpdb.cgi**

7. **Now we will choose and supply the template structure**. Get the PDB file **1iep** and save it locally on your computer. The program has also an option to choose the template files for you. We will not use this option now. Open the file by Swiss PDB viewer.

8. Choose "**Alignment**" from the "Window" menu.

9. Click on the **pka** name to make this layer active. Choose the "**Magic Fit**" option of the "**Fit**" menu. This will perform the sequence alignment. Choosing "**Improved Fit**" from the same menu will optimize the alignment.

10. Make sure all residues of the 2 proteins are selected. From the "Color" menu choose "color by alignment diversity", so you will be able to identify the conserved regions. 11. Choose the "**Update threading now**" item of the

"SwissModel" menu (this item is not accessible if the "Update Threading Display automatically" item is enabled; which is the case by default).

12. After the initial automatic alignment we have now the freedom to change it. This is done with the mouse and the arrow keys. We can also make use of the mean force potential to help threading correctly a protein, although this tool should be used with caution. Make sure the current layer is pka, and click on the little arrow located at the right of the question mark of the Alignment Window. The window expands, and displays a curve depicting how each residue likes it's surrounding. If a residue is "happy", its energy is below zero, whereas unhappy residues will have energy above the zero axis. This is the mean force potential energy. Click on the "smooth" text, and set a smoothing factor of 1. It means that the energy of each residue will be the average of itself plus the energy of 1 flanking residue on each side. You can enable the "Auto Color by Threading Energy" item of the "SwissModel" menu to better see the potential on the structure. Click on the "E= XX" text, this will re-compute the energy for the current layer. **Note:** this tool provides hints and should be used in conjunction with other type of analyses! It works better for displacement of large fragments.

13. You can also evaluate how good your threading is by using the "aa making clashes" items of the "Select" menu. This will allow you to quickly focus on potentially problematic regions. You can then choose the "Fix Selected Side-chains" ("crude") item of the "Tools" menu, which will browse the rotamer library to choose the best rotamer, exactly as during a mutation process. By repeating the "Select aa making clashes" process, you should see that fewer amino-acids are making problems. If not, this is probably a good clue that your threading is incorrect. **Important Note:** Actually fixing the side-chains is just for you, to evaluate prior to submitting the request, it will have absolutely no incidence onto the model building, as side-chains are reconstructed anyway.

14. When everything seems OK, you can submit a modeling request to Swiss-Model simply by choosing the "Submit modeling request" of the "SwissModel" menu. You will be asked to give a project name. By the default, you will get to your email a Swiss-PDB-Viewer project file, with your model aligned onto the templates you used, and ready for comparison.

15. While the server is working, we will compare the results of the first approach mode with the real structure for that protein. **Open the structural alignment program CE** (combinatorial extension) at the URL: **http://cl.sdsc.edu/ce/ce_align.html**. For the first chain upload the model you obtained from Swiss-Model. Don't forget to mark the "User File" option instead

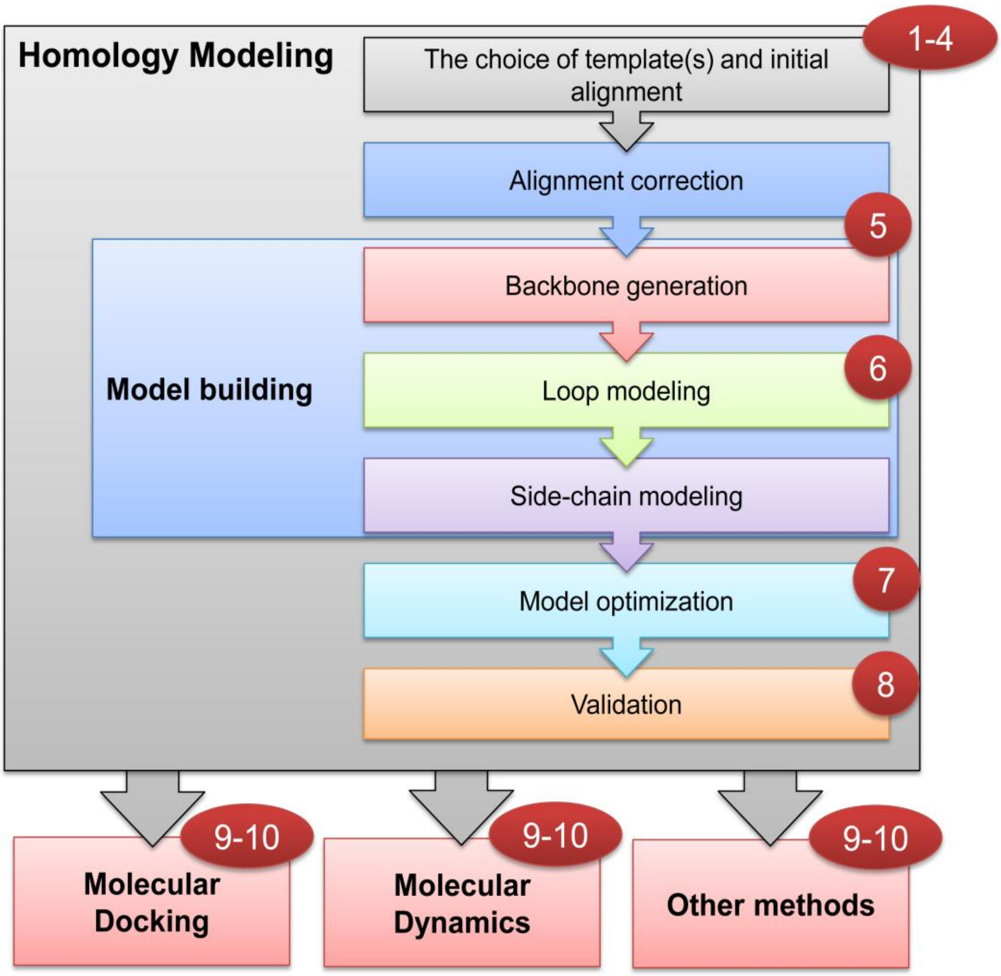
of the “PDB” option. For the second chain, enter 1APM:E which are the PDB code and chain identifier for the real structure exists for PKA. Submit.

16. At the results page, look at the alignment. Notice that this sequence alignment was produced according to the structural information, without considering the sequence. It allows us to judge the structural similarity. Find the regions which were not properly built. What is the overall RMSD of the structural alignment? Is this significant?

17. Save the PDB format of the structural alignment in your computer and open it with **RasMol**. Find the regions not properly aligned by visual inspection and compare to your answer from the previous question.

18. Finally we will take a look at the at the **evaluation** report for this model. Look at the evaluation graphs obtained using Anolea, Gromos and Verify3d**.** Try to find regions which are suspected to be incorrect.

19. To conclude: structural model is basically easy to obtain, but we always be aware of how it was produced, check it with available tools and refine it if necessary.



KEGG (**Kyoto Encyclopedia of Genes and Genomes**) is a collection of online databases dealing with genomes, enzymatic pathways, and biological chemicals. The PATHWAY database records networks of molecular interactions in the cells, and variants of them specific to particular organisms

KEGG maintains five main databases

- KEGG Atlas
- KEGG Pathway
- KEGG Genes
- KEGG Ligand
- KEGG BRITE

Metabolism is the cellular subsystem responsible for generation of energy from nutrients and production of building blocks for larger macromolecules. Computational and statistical modeling of metabolism is vital to many disciplines including bioengineering, the study of diseases, drug target identification, and understanding the evolution of metabolism.

Kegg pathway is a collection of manually drawn pathway maps (see new maps and update history) representing our knowledge on the molecular interaction and reaction networks for:

1. Metabolism

Global map, Carbohydrate, Energy, Lipid, Nucleotide, Amino acid, Other amino acid

Glycan, Cofactor/vitamin, Terpenoid/PK, Other secondary metabolite, Xenobiotics

Reaction module, Chemical structure etc..

2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases

and also on the structure relationships (KEGG drug structure maps) in:

7. Drug Development

KEGG **Pathway Mapping** is the process to map molecular datasets, especially large-scale datasets in genomics, transcriptomics, proteomics, and metabolomics, to the KEGG pathway maps for biological interpretaion of higher-level systemic functions.

•Search Pathway - basic pathway mapping tool
•Search&Color Pathway - advanced pathway mapping tool
•Color Pathway - selected pathway map coloring tool

The contents of KEGG include the following topics

- ▼ **Metabolism**
  - ▼ Global map
    - Metabolic pathways
    - Biosynthesis of secondary metabolites
    - Microbial metabolism in diverse environments
  - ▶ Carbohydrate metabolism
  - ▶ Energy metabolism
  - ▶ Lipid metabolism
  - ▶ Nucleotide metabolism
  - ▶ Amino acid metabolism
  - ▶ Metabolism of other amino acids
  - ▶ Glycan biosynthesis and metabolism
  - ▶ Metabolism of cofactors and vitamins
  - ▶ Metabolism of terpenoids and polyketides
  - ▶ Biosynthesis of other secondary metabolites
  - ▶ Xenobiotics biodegradation and metabolism
  - ▶ Reaction module maps
  - ▶ Chemical structure transformation maps

- ▼ **Genetic Information Processing**
  - ▶ Transcription
  - ▶ Translation
  - ▶ Folding, sorting and degradation
  - ▶ Replication and repair
- ▼ Environmental Information Processing
  - ▶ Membrane transport
  - ▶ Signal transduction
  - ▶ Signaling molecules and interaction

- ▼ Cellular Processes
  - ▶ Transport and catabolism
  - ▶ Cell motility
  - ▶ Cell growth and death
  - ▶ Cell communication

- ▼ Organismal Systems
  - ▶ Immune system
  - ▶ Endocrine system
  - ▶ Circulatory system
  - ▶ Digestive system
  - ▶ Excretory system
  - ▶ Nervous system
  - ▶ Sensory system

  - ▶ Development
  - ▶ Environmental adaptation

- ▼ Human Diseases
  - ▶ Cancers: Overview
  - ▶ Cancers: Specific types
  - ▶ Immune diseases
  - ▶ Neurodegenerative diseases
  - ▶ Substance dependence
  - ▶ Cardiovascular diseases
  - ▶ Endocrine and metabolic diseases
  - ▶ Infectious diseases: Bacterial
  - ▶ Infectious diseases: Viral
  - ▶ Infectious diseases: Parasitic

- ▼ Drug Development
  - ▶ Chronology: Antiinfectives
  - ▶ Chronology: Antineoplastics
  - ▶ Chronology: Nervous system agents
  - ▶ Chronology: Other drugs
  - ▶ Target-based classification: G protein-coupled receptors
  - ▶ Target-based classification: Nuclear receptors
  - ▶ Target-based classification: Ion channels
  - ▶ Target-based classification: Transporters
  - ▶ Target-based classification: Enzymes
  - ▶ Structure-based classification
  - ▶ Skeleton-based classification

Over view of SWISS PROT

SWISS-PROT is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the European Molecular Biology Laboratory (EMBL), since 1987. SWISS-PROT is now an equal partnership between the EMBL and the Swiss Institute of Bioinformatics (SIB). The EMBL activities are carried out by its Hinxton Outstation, the European Bioinformatics Institute (EBI)..

The SWISS-PROT protein sequence database consists of sequence entries. Sequence entries are composed of different line types, each with their own format. For standardisation purposes the format of SWISS-PROT (see http://www.expasy. ch/txt/userman.txt) follows as closely as possible that of the EMBL Nucleotide Sequence Database. A sample SWISS-PROT entry is shown in http://www.expasy.ch/cgi-bin/niceprot. pl?P29965

The SWISS-PROT database distinguishes itself from other protein sequence databases by three distinct criteria: (i) annotations, (ii) minimal redundancy and (iii) integration with other databases.

Annotation

In SWISS-PROT two classes of data can be distinguished: the core data and the annotation. For each sequence entry the core data consists of the sequence data; the citation information (bibliographical references) and the taxonomic data (description of the biological source of the protein), while the annotation consists of the description of the following items:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, homeoboxes, SH2 and SH3 domains, etc.
- Secondary structure. For example alpha helix, beta sheet, etc.
- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiencie(s) in the protein
- Sequence conflicts, variants, etc.

We try to include as much annotation information as possible in SWISS-PROT. To obtain this information we use, in addition to the publications reporting new sequence data, review articles to periodically update the annotations of families or groups of proteins. We also make use of external experts who have been recruited to send us their comments and updates concerning specific groups of proteins (see http://www.expasy. ch/cgi-bin/experts).

We believe that the systematic recourse both to publications other than those reporting the core data and to subject referees represents a unique and beneficial feature of SWISS-PROT. In SWISS-PROT, annotation is mainly found in the comment lines (CC), in the feature table (FT) and in the keyword lines (KW). Most comments are classified by ‘topics’; this approach permits the easy retrieval of specific categories of data from the database.

SWISS-PROT is a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recent developments of the database include format and content enhancements, cross-references to additional databases, new documentation files and improvements to TrEMBL, a computer-annotated supplement to SWISS-PROT. TrEMBL consists of entries in SWISS-PROT-like format derived from the translation of all coding sequences (CDSs) in the EMBL Nucleotide Sequence Database, except the CDSs already included in SWISS-PROT. We also describe the Human Proteomics Initiative (HPI), a major project to annotate all known human sequences according to the quality standards of SWISS-PROT. SWISS-PROT is available at: http://www.expasy.ch/sprot/ and http://www.ebi.ac.uk/swissprot/

Computer-Aided Drug Design (CADD)

CADD helps scientists in minimizing the synthetic and biological testing efforts by focussing only on the most promising compounds. Besides explaining the molecular basis of therapeutic activity, it also predicts possible derivatives that would improve activity. CADDD entails (Kapetanovic, 2008):

(1) Drug discovery and development processes being streamlined by the use of computing power.

(2)Identification and optimization of new drugs using leverage of chemical and biological information about targets and/or ligands.

(3)In silico designing of filters for the elimination of undesirable compounds with properties like poor activity and/or poor absorption, distribution, metabolism, excretion and toxicity, ADMET which facilitate selection of the most promising candidates.

Advantages of CADD

The main advantages of drug discovery through CADD are:

(i)For experimental testing, smaller set of compounds are selected from large compound libraries.

(ii)Drug metabolism and pharmacokinetics (DMPK) properties like absorption, distribution, metabolism, excretion and the potential for toxicity (ADMET) are increased by optimization of lead compounds.

(iii)Designing of novel compounds can be achieved either by “growing” starting molecules one functional group at a time or by piecing together fragments into novel chemotypes (Veselovsky and Ivanov, 2003).

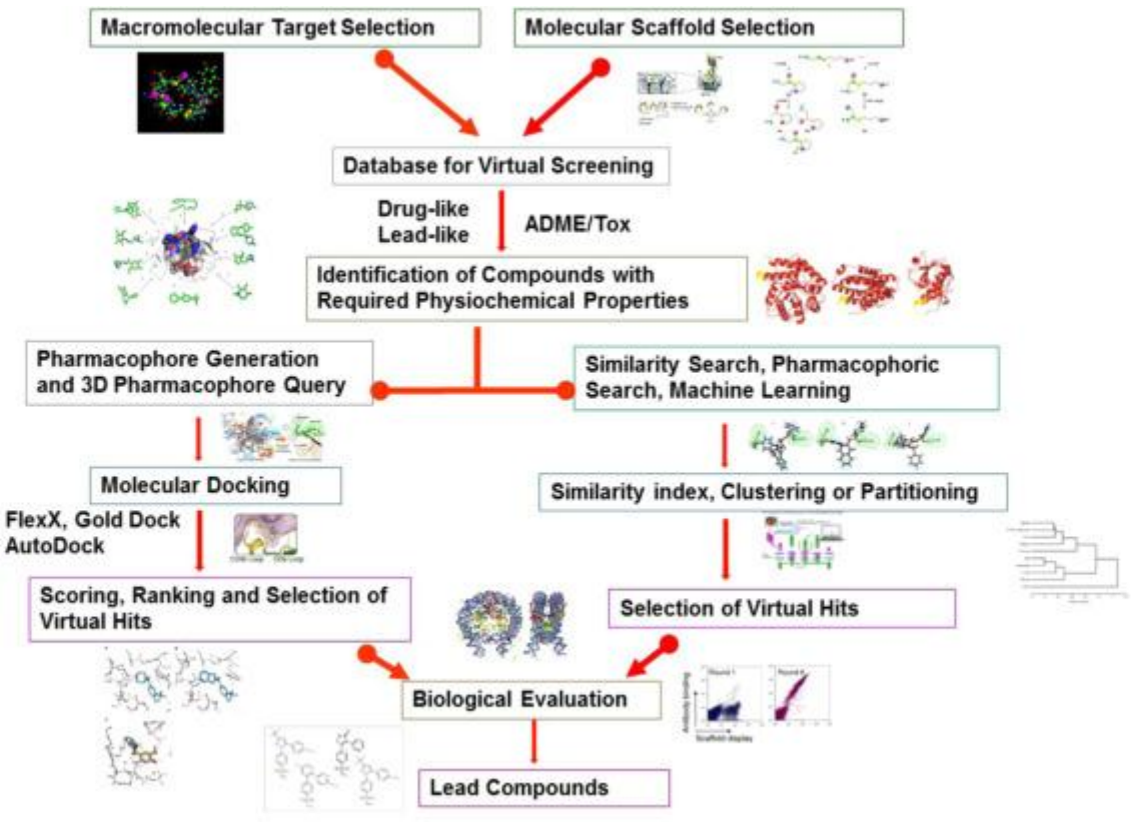
(iv)Traditional experimentation which requires animal and human models can be replaced by CADD, saving both time and cost.

(v)Reduces the chances of drug resistance and thus would lead to production of lead compounds which would target the causative factor.

(vi)CADD also leads to the construction of high quality datasets and libraries that can be optimized for high molecular diversity or similarity.

Types of CADD

The choice of CADD approaches to be employed is determined by the availability of the experimentally determined 3D structures of target proteins. Structure-based CADD uses our knowledge of the target protein structure to calculate interaction energies, whereas in ligand-based CADD, chemical similarity searches or construction of predictive, quantitative structure-activity relationship (QSAR) models exploits our knowledge of known active and inactive molecules.(Kalyaanamoorthy and Chen, 2011). Structure based CADD combines information from several fields, for example, X-ray crystallography and/or NMR, synthetic organic chemistry, molecular modelling, QSAR, and biological evaluation. Through structure based CADD, we aim to design compounds with strong binding affinity with the target, thereby exhibiting properties like reduced free energy, improved DMPK/ADMET properties and target specification i.e., reduced off-target. Virtual high-throughput screening (vHTS) also known as screening of virtual compound libraries is one of the most common applications of CADD Fig. 6 represents an overview of CADD drug designing/design pipeline.

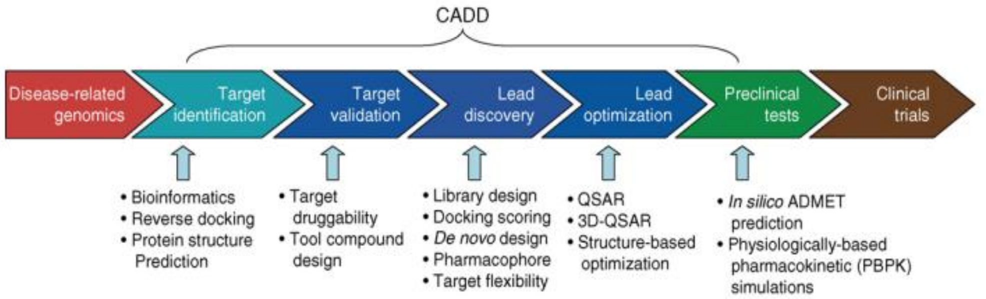

Macromolecular Target Selection
Molecular Scaffold Selection
Database for Virtual Screening
Drug-like
Lead-like
ADME/Tox
Identification of Compounds with Required Physiochemical Properties
Pharmacophore Generation and 3D Pharmacophore Query
Similarity Search, Pharmacophoric Search, Machine Learning
Molecular Docking
Similarity index, Clustering or Partitioning
FlexX, Gold Dock
AutoDock
Scoring, Ranking and Selection of Virtual Hits
Selection of Virtual Hits
Biological Evaluation
Lead Compounds

An overview of CADD drug designing/design pipeline.

Adapted and modified from Guido, R.V., Glaucius Oliva, G., Andricopulo, A.D., 2008. Virtual screening and its integration with modern drug design technologies. Current Medicinal Chemistry 15, 37–46.

Structure–based drug discovery

This method exploits knowledge of the three-dimensional structure of a receptor complexed with a lead molecule for optimization of the bound ligand or a series of congeneric molecules. It requires the understanding of receptor–ligand interactions. The structural information can be obtained either from X-ray crystallography, NMR, or from homology modelling. A medicinal chemist can use a model with a given structure for computing the activity of a molecule (Lewis, 2005). Some of these approaches provide accurate binding modes, while cater to fast searching of large databases. Some approaches of structure-based drug designing are explained below.



Drug Discovery Today: Technologies

Figure 1. Drug discovery pipeline vs. computer-aided drug design (CADD) tools. CADD tools have been initially developed for lead optimization (such as QSAR) and then expanded for lead discovery (such as virtual screening). Now CADD tools have quickly extended toward both upstream and downstream directions along the drug discovery pipeline. In upstream direction, bioinformatics and reverse docking methods are usually used for target identification; once a target is identified, *in silico* methods are also developed to predict its 3D structures before experimental determination; computational methods can be used to predict target druggability and to design compounds before further experiments. In downstream direction, *in silico* ADMET prediction and physiologically based pharmacokinetic simulations can be conducted to model the preclinical test, which is usually integrated into *in silico* lead discovery stage to reduce the costs.

Structure-based virtual high-throughput screening

Structure-based virtual high-throughput screening (SB-vHTS) is an in-silico method which helps identify putative hits out of hundreds of thousands of compounds to the targets of known structure. It is usually based on molecular docking. In molecular docking, a small molecule is fitted into the active site of protein model and here, comparison of the 3D structure of small molecule with the putative binding pocket is carried out. In the traditional HTS, the general ability of a ligand to bind, inhibit or allosterically alter the proteins function is asserted experimentally, whereas in SB-vHTS selects the ligands that are predicted to bind to a specific binding site. To ensure the feasibility of screening of large compound libraries within a finite time, limited conformational sampling of proteins and ligands is used by SB-vHTS along with a simplified approximation of binding energy that can be computed rapidly (Becker et al., 2006

Structure-based virtual screening

This is a computational approach for identifying potential drug candidates (hits) that are capable of binding to a drug target (protein receptors, enzymes). This method involves quick searching of large libraries of chemical followed by docking of the hit into a protein target and finally application of a scoring function for estimating the probability of binding affinity of drug candidate with the protein target (Cheng et al., 2012). The most important advantage of this screening is that it enhances the hit rate by considerably decreasing the number of compounds that are estimated experimentally for their activity and hence improves the success rate of the in vitro experiments. This method has been applied extensively in pharmaceutical companies and academic groups for early-stage drug discovery.

Fragment-based lead discovery

This approach is based upon structure-activity relationships (SAR), obtained from NMR for identifying and optimizing the lead (Bienstock, 2011). High purity, weak potency but effective binding, good aqueous solubility, (molecular weight<300, ClogP<3, number of rotatable bonds, number of hydrogen bond donors and acceptors each should be <3) are the criteria for selecting the chemical fragments (Congreve et al., 2003). Later, these fragments are either expanded or combined for producing a lead with a higher affinity.

In silico structure-based lead optimization

After the desired hits are identified through virtual screening, this method speeds up the search for optimized lead by delineating the prediction about its pharmacological properties, thereby reducing the in vitro and in vivo experimental time.

ADMET modelling

This method, a common name for which is physiologically-based pharmacokinetic modelling is used in drug design and development, and in assessing of toxicity threat evaluation and specifically predicts absorption, distribution, metabolism, excretion and toxicology (ADMET) of drugs/compounds in humans. The ADMET parameters are based on the kinetics of the drug exposure to tissues and how the body will react to them, influencing the performance and pharmacological activity of the compound. Therefore, this method provides a key insight into the behaviour of a pharmaceutical compound within an organism. This approach aids in the selection of compounds during the very early phases of drug thereby playing a crucial role in drug discovery and development. This technique is cost- and time effective owing to a reduction in attrition of drugs during the pre-clinical / clinical phase trials at a later stage.

Ligand-based drug designing

The existing knowledge of active compounds against the target is used to predict new chemical entities that present similar behaviour in Ligand-based methods (Martin et al., 2002). Given a single known active molecule, a pharmacophore model can be derived from a library of molecules to define the minimum necessary structural characteristics a molecule must possess in order to bind to the target of interest. A fingerprint-based similarity search is usually used to compare the active molecule to the library as here, the molecules are represented as bit strings which represent the presence or absence of predefined structural descriptors (Mishra and Siva-Prasad, 2011). In comparison, targeting structural information to determine whether a new compound is likely to bind and interact with a receptor is the method that structure-based methods rely on. No prior knowledge of active ligands is required in this method, which is a significant advantage (Kolb et al., 2009). It is possible to design new ligands that can elicit a therapeutic effect from 3D structures. Therefore, the development of new drugs through the discovery and optimization of the initial lead compound are greatly impacted by structure-based approaches.

Ligand-based virtual screening (LBVS)

Ligand-based virtual screening is based on the "similarity principle" according to which similar molecules tend to exhibit similar biological properties. Scaffold hopping i.e., identification of iso-functional molecular structures with significantly different molecular backbones is the usual objective when using LBVS. "Scaffold hopping" is also known as "leapfrogging", "scaffold searching" and "leap hopping" (Kalliokoki, 2010). These methods are usually helpful in drug repurposing, wherein new targets and diseases are pursued for existing drug molecules.

Molecular descriptors

This is one of the simplest approaches in which the reference molecule/set of molecules are compared with a large library of compounds at a very low cost on the basis of physicochemical properties descriptors, such as molecular weight, volume, geometry, surface areas, atom types, dipole moment, polarizability, molar refractivity, octanol-water partition coefficient (log P), planar structures, electronegativity, or solvation properties that are obtained from experimental measurements or theoretical models. Molecules are represented by symbols for effective execution of the task (Prada-Graciaa et al., 2016).

Quantitative structure-activity relationship models (QSAR)

The mathematical relation between structural attributes and target response for a set of chemicals are explained by Quantitative Structure-Activity Relationship models. Structural and/or property descriptors of compounds can also be correlated with their biological activities using QSAR (Bernard et al., 2005). Through QSAR models we can correlate various features like rate constants, binding sites affinities of ligands, inhibition constants and other biological activities, either with certain structural features (Free Wilson analysis) or with atomic, molecular or group properties, such as lipophilicity, electronic, steric and polarizability, among congeneric series of compounds. (Kubinyi, 1995). Hence, the success of QSAR is dependent on the choice of descriptors and the ability to generate the appropriate mathematical relationship besides the quality of initial set of active/inactive compounds.

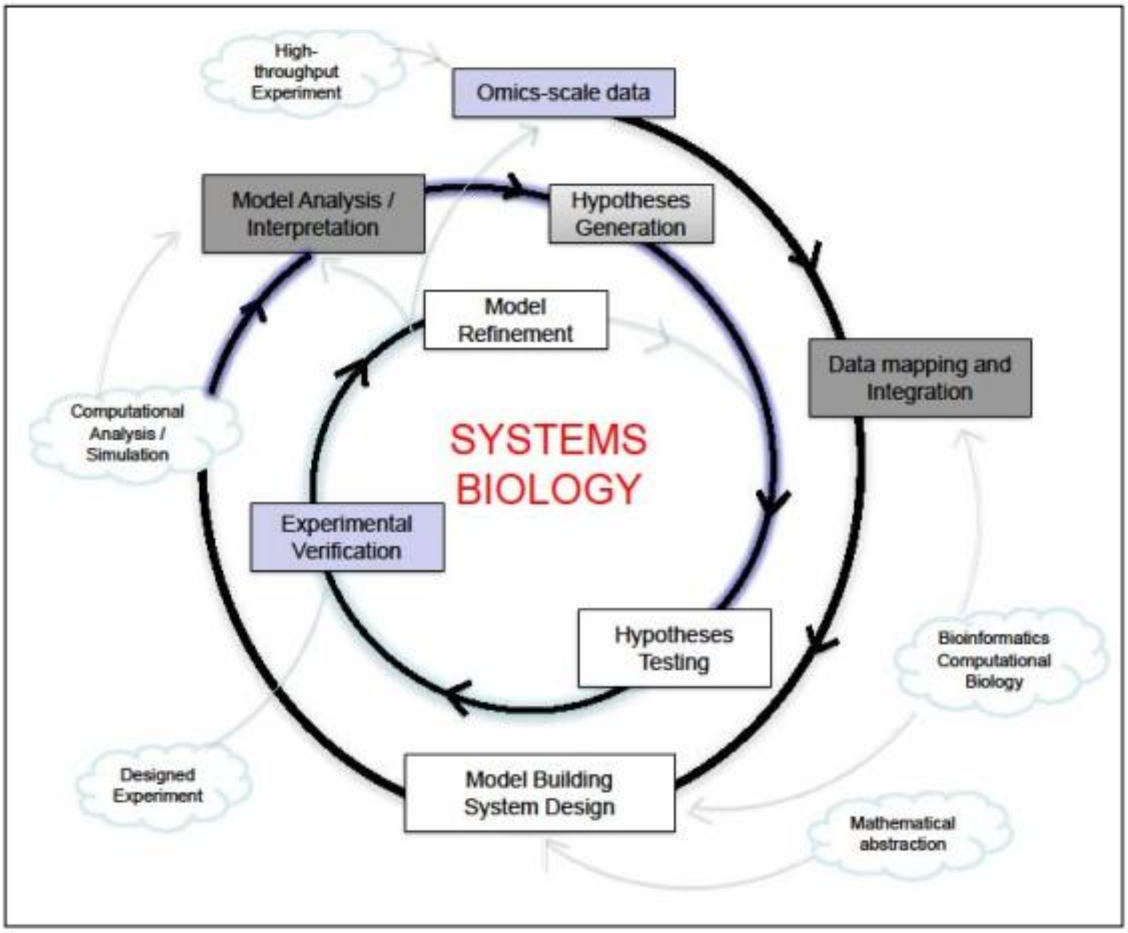
Pharmacophore modelling

More significant information can be drawn by employing various conformations of a range of ligands than just a single ligand structure. A pharmacophore model of the receptor site can be generated with a sufficiently broad range of ligands. Pharmacophore modelling of smaller, non-peptide molecules that might have improved stability and bioavailability over their peptide counterparts has resulted in successful outcomes so far (Nielsen et al., 1999).

Systems biology

Systems biology seeks to study biological systems as a whole, contrary to the reductionist approach that has dominated biology. Such a view of biological systems emanating from strong foundations of molecular level understanding of the individual components in terms of their form, function and interactions is promising to transform the level at which we understand biology. Systems are defined and abstracted at different levels, which are simulated and analysed using different types of mathematical and computational techniques. Insights obtained from systems level studies readily lend to their use in several applications in biotechnology and drug discovery, making it even more important to study systems as a whole.

Systems biology, being a holistic approach involves modelling and analysis of metabolic pathways, regulatory and signal transduction networks for understanding cellular behaviour. There are also various levels of abstraction at which these systems are modelled, with a wide variety of techniques that can be employed based on the quality and quantity of data available.


High-throughput Experiment
Omics-scale data
Model Analysis / Interpretation
Hypotheses Generation
Model Refinement
Data mapping and Integration
Computational Analysis / Simulation
SYSTEMS BIOLOGY
Experimental Verification
Hypotheses Testing
Bioinformatics Computational Biology
Designed Experiment
Model Building System Design
Mathematical abstraction

Systems biology process. This process relies on an iterative procedure of model building, experimental verification, model analysis and model refinement. The concepts that underlie these processes have been shown as clouds.