Aims of this Laboratory:

For the current lab, you will explore several aspects of structural analyses. Structures of macromolecules provide insight into protein function and interactions with solvent, ligands and other macromolecules. In these studies, we will explore several different approaches to structural analyses. The following introductory material describes theoretical aspects of structural studies, as well as, examples of how to set up various types of structural analysis experiments involving mass spectrometry, x-ray crystallography, and computational approaches to structural questions.

EXPERIMENT 1

Mass spectrometry (MS) is the measure of mass-to-charge ratio as well as the abundance of gas-phase ions. This technique provides a very precise measure of a molecule's molecular weight. Additionally, MS provides information on a molecule's chemical structure through the degradation products resulting from a molecule's collisions with electrons. Accurate measurement of a molecule or a fragment of molecule allows a means to identify unique components of complex mixtures, like cellular lysates, macromolecular complexes or intermediates of reactions.

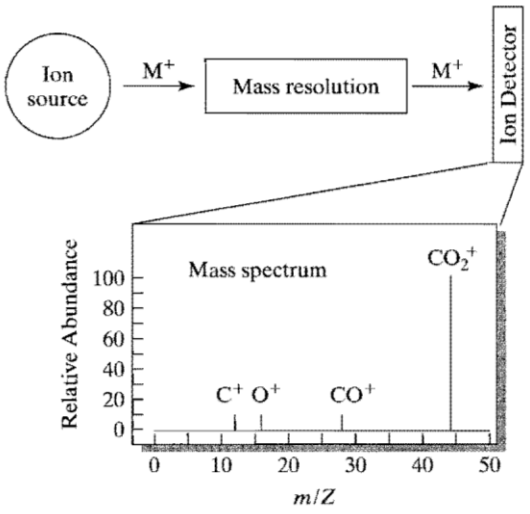


Figure 15.1 The three primary components of a mass spectrometer. The ion source generates molecular ions M^+ that can then be resolved (separated or filtered) according to their mass to charge ratios (m/Z) for detection. The mass spectrum of CO_2 after electron ionization is shown.

Figure 1.1 Overview of MS experiment. Taken from *Principles of Physical Biochemistry*, Van Holde.

An MS experiment consists of three components: 1) The molecule of interest is converted to a charged ion in the gas phase; 2) The molecule is separated by mass; and 3) The molecular ions are detected following separation. Separation is achieved via the electrostatic potential of the molecular ions. See Figure 1.1. Molecular ions of small mass or larger charge (mass-to-charge (m/Z) ratio) have a higher velocity, providing a means to separate ions based upon m/Z differences.

Two approaches have been developed for MS of macromolecules: MALDI (Matrix Assisted Laser Desorption Ionization) and ESI (ElectroSpray Ionization). For this lab, we will be completing and ESI experiment. In ESI, a charged microdroplet, consisting of the macromolecule(s) to be examined, is sprayed into the mass spectrometer through a charged nozzle. Ionization of the droplet occurs as the sample exits the nozzle. As the droplets accelerate away from the tip, the solvent evaporates until the Coulombic forces overcome the surface tension of the drop resulting in dispersion of the drop into a spray of smaller droplets. This process continues until all solvent is removed leaving the macroions for analyses, Figure 1.2. Macroions can be singly or multiply charged (m/Z where $Z \geq 1$).

The accurate determination of a molecule molecular weight can be used to identify a macromolecule within a complex mixture. A typical approach to identifying components of a complex mixture, such as macromolecular complexes, is to digest the mixture with a site-specific protease. Trypsin, which cleaves a peptide bond following an arginine or lysine

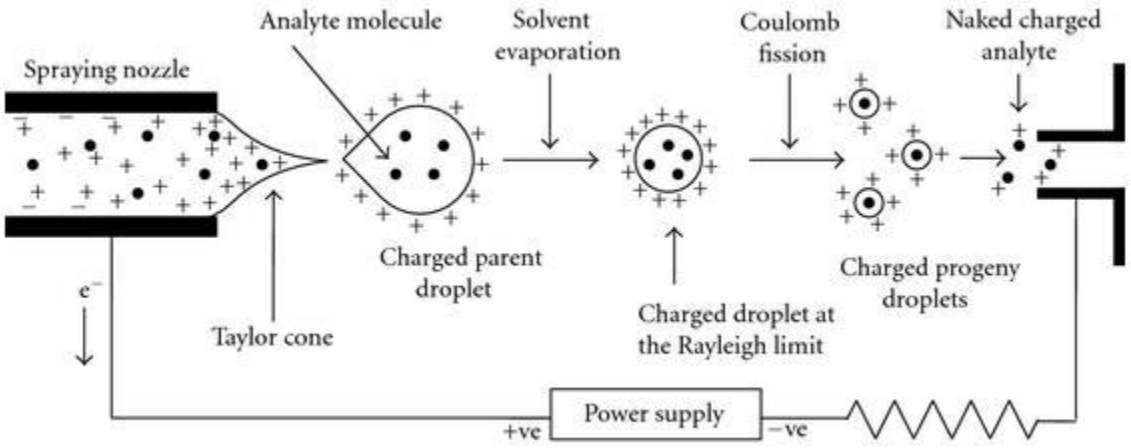


Figure 1.2. Schematic of ESI. Adapted from Intl. J. Anal. Chem. 2012 Art. ID 282574.

residue, is the most commonly used protease. Other commonly used, site-specific proteases are listed in Table 1.1. The digested mixture is then separated via high pressure liquid chromatography (HPLC) on a C12 or C18 column using a formic acid:acetonitrile gradient (0.1% formic acid to 0.1% formic acid, 100% acetonitrile). The eluted peaks are fed directly into the MS nozzle for analyses. A single protein digested with a site-specific protease will give a unique, predictable cleavage pattern, a protein fingerprint, that can be detected by MS. Yet, in a complex mixture, peptide fragments may overlap, be undetected, or unique fragments may be missing from the sample due to missed cleavage sites. An experimenter will generate a list of cleavage sites and predicted fragments that can then be compared to the experimental data to calculate a probability that the cleavage pattern observed in the experimental data is a "match" to target protein. Several web-based tools are available to generate a list of peptides generated from a protease digestion. In this lab, we will use Protein Prospector (http://prospector.ucsf.edu/prospector/mshome.htm) to predict the cleavage pattern of porcine mitochondrial malate dehydrogenase.

To uniquely identify the protease-generated peptides, a technique called tandem mass spectrometry or MS/MS is used. In this process, the mass of a peptide is measured by MS and then ions of a defined mass are fed into a second chamber, a collision chamber. The collision chamber is filled with a neutral gas (argon or xenon) that, upon colliding with the peptide, breaks the peptide backbone. This process is called collision-induced dissociation and results in daughter ions that can be fed into a second mass analyzer. See Figure 1.3 for an overview of this process. From these analyses, unique peptide fragments can be identified and the identity of a protein within a mixture confirmed.

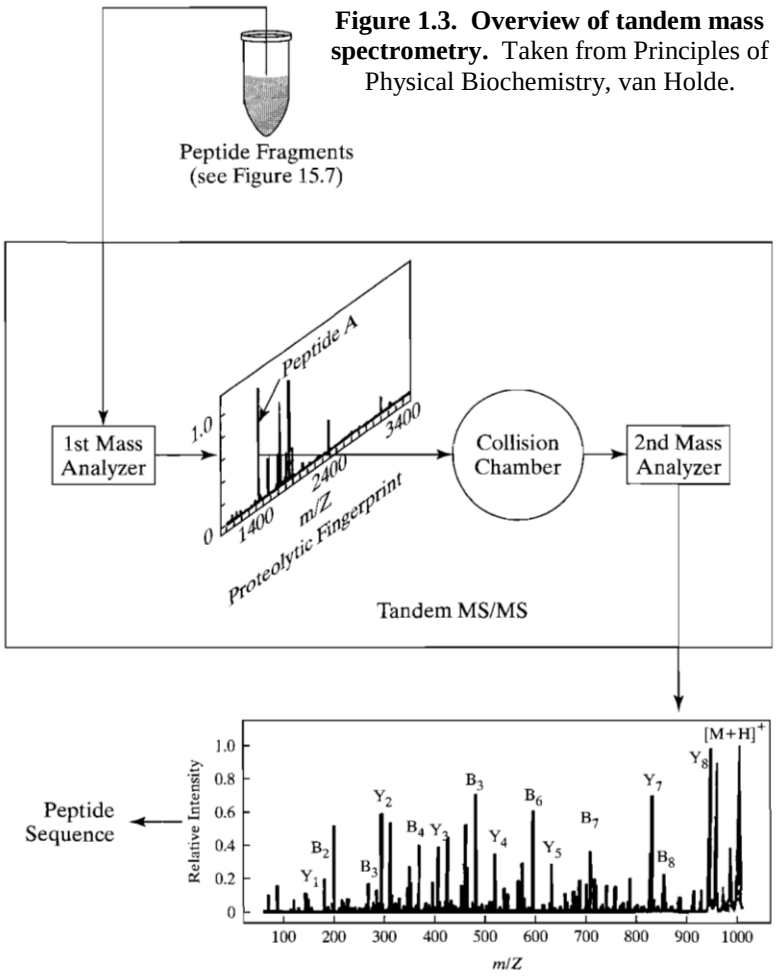


Figure 1.3. Overview of tandem mass spectrometry. Taken from Principles of Physical Biochemistry, van Holde.

Figure 15.16 Peptide sequencing by tandem mass spectrometry (MS/MS). The sequence of a mixture of proteolytic peptide fragments is separated by the first mass spectrometer (in a manner similar to producing a peptide fingerprint, see Figure 15.12). One peptide (labeled Peptide A) is specifically sent to a collision chamber for further fragmentation into daughter ions, and the daughter ions are analyzed by the second mass spectrometer. The m/Z for the daughter ions (labeled as the B- and Y-ions) are then used to reconstruct the sequence of the Peptide A.

Table 1.1. Common Proteolytic Enzymes.

Enzyme	***Preference @ R1***	***Preference @ R2***
Trypsin	K, R	≠ P
Chymotrypsin	T, F, L, I, V, W, His @ high pH	≠ P
Pepsin	F, L	≠ P
Thrombin	R	≠ P
Papain	R, K, F	none
V8	E, Q (pH = 4)	None
Thermolysin	≠ P	T, F, L, I, V, W, His @ high pH

Table 1.2. Residue masses of 20 common amino acids.

Name	***1-letter code***	***Residue mass (Da)***
Glycine	G	57.02
Alanine	A	71.04
Serine	S	87.03
Proline	P	97.05
Valine	V	99.07
Threonine	T	101.05
Cysteine	C	103.01
Isoleucine	I	113.08
Leucine	L	113.08
asparagine	N	114.04
Aspartic Acid	D	115.03
Glutamine	Q	128.06
Lysine	K	128.09
Glutamic acid	E	129.04
Methionine	M	131.04
Histidine	H	137.06
Phenylalanine	F	147.07
Arginine	R	156.1
Tyrosine	Y	163.06
Tryptophan	W	186.08

Note: The masses of the isolated amino acids are the residue massess plus 18 Da.

Limitations of this approach arise from pairs of amino acids that simply cannot be distinguished, like leucine and isoleucine (113.08 Da). Table 1.2 gives the residue masses of the 20 common amino acids. In addition, one must understand the generation of daughter ions in order to determine the sequence. The peptide backbone consists of the C_α-C, C-N, and N-C_α bonds and each bond can be broken via collision to generate fragments of the parent peptide, see Table 1.3. Note, breaking any of the three bonds generates two complementary fragments, one representing the mass of the amino acids toward the N-terminus of the break and a second representing the mass toward the C-terminus. The daughter ions are labeled A, B, and C for the N-terminal products resulting from fragmentation at the C_α-C, C-N, and N-C_α, respectively. The C-terminal fragments are X, Y, Z and complement A, B, and C, respectively. The B- and Y-ions are the most commonly observed pair of complementary daughter ions.

The mass of a peptide (S) is the sum of the amino acid residue masses as given in Table 1.2, which are 18 Da less than the natural isolated residues to account for the water lost in forming the peptide bond, plus 1 for the hydrogen at the N-terminus and 17 for the OH at the C-terminus of the parent peptide fragment. The total mass (m) for a parent peptide is:

$$m = \sum_{i=1}^{n} S_i + 18$$

The parent peptide is (m + 1)/1 @ Z = 1; (m + 2)/2 @ Z = 2; etc. Given the mechanism in Figure 1.4 for generating B- and Y- ions, one can see that a “true” B- ion is not observed for the n=1 cleavage. The B-ions are expected to generate a set of peaks in the spectrum that can be read from left to right, and the Y-ions peaks that can be read from right to left. An example of the assignments for B- and Y- ions is given for a short peptide (reproduced from Principles of Physical Biochemistry, van Holde.) in Figure 1.5.

Experiment 1.1. Identification of a porcine mMDH via MS/MS analyses

The instrument is extremely sensitive. Do NOT handle reagents without gloves. You will deposit keratin into your sample (constantly shed in the form of skin cells) and it will contaminate your sample/ complicate your results. Also, do NOT wear any wool clothing in the ESI-MS instrument room as it also will deposit keratin in the instrument.

Reagents:

Figure 15.18 Proposed mechanism for generation of B- and Y-ions from CID fragmentation. The B-ion is generated directly from the proposed cleavage reaction, while the Y-ion requires a proton transfer. For this reason, there is a higher possibility for the exchange of the proton with protons from solvent in the Y-ions as compared to the B-ions.

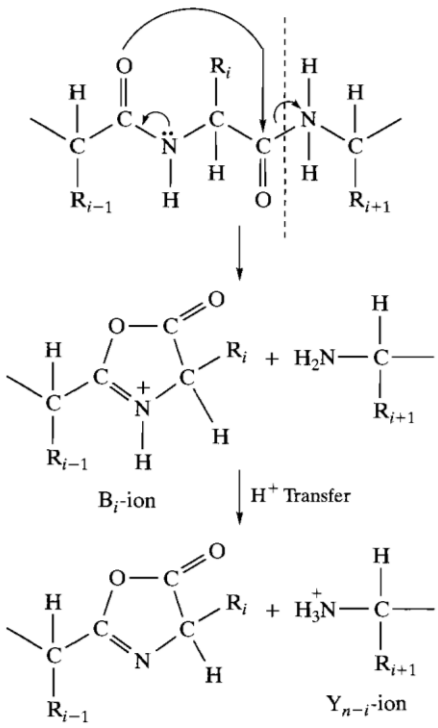


Figure 1.4. Mechanism for generating B- and Y-ions. Taken from Principles of Physical Biochemistry, van Holde.

Reduced, alkylated and trypsin digested porcine mitochondrial malate dehydrogenase

Materials/Equipment:
HPLC equipped with C12 column
LTQ – ESI (ESI-TRAP)

Step-by-step instructions for MS/MS analysis:
The instrumentation consists of three major components: Liquid and sampler handler; chromatography holder & diode array; and electrospray ionization mass spectrometer.

Start-up procedure:
The U3400RS (liquid/sample handler & chromatography holder) and ESI-MS should have been left in stand-by mode. The buttons on the front of the instrumentation will be lit red.
Bring the instruments online by pushing each of the buttons. The lights will turn green and the instrument will go through a self-check.

Check the volumes of your reagents in lines A and B:
A: 0.1% formic acid in water
B: 0.1% formic acid in acetonitrile
C: 100% acetonitrile

You need a minimum of 200 ml per reagent before starting the run. If insufficient reagents available, ask your instructor for assistance in replenishing these reagents.

Load sample vial into the sample holder. Note placement of sample in terms of tray and well.
TRAY:_____ (color)
WELL: _____ (Letter/#)

Turn on N_2 tank (2 turns to the left)

On MS, remove needle from nozzle.

On the computer, open Chromeleon Xpress.
Under system, click on "Take control."
Click on tab "HPG-3400RS" (pump console).
Click "Connect."
Choose flow rate **0.07ml/min**.

Table 1.3. Characteristics of CID daughter ions.

ion type	*composition*	*change in mass*
A	+H, -CO	-27
B	+H	+1
C	+3H, +NH	+18
X	+OH, +CO	+45
Y	+OH, +2H	+19
Z	+OH, -NH	-27

Composition: # & types of atoms added to peptide fragment to generate daughter ion

Change in mass is added (or subtracted) mass to the sum of the residue masses of the amino acids in a peptide fragment to calculate the m/Z ratio of the daughter ion

Choose mobile phase composition: A = 95% (0.1% formic acid in water); B = 5% (0.1% formic acid in acetonitrile)
You are now equilibrating the column.
Reduce the Chromeleon software to the software bar.

On the computer, open Xcalibur.
The status window (4th button from left gives status bar if not open) should show that both the chromatography system and ESI-MS are "Ready to Download."
Click on the "Instrument setup" icon.
Choose File >Open > File name = CHEM427….meth (C:\xcalibur\methods\)
In the method file:
In the left column: Dionex Chrom.
- Controls LC
- Control column oven (30°C)
- Pump
  - Gradient type (Do NOT change, FYI: 5 = linear)
- Sampler
  - Prep for sample
  - Temperature for sample compartment (usually 20°C)

In the left column: LTQ XL/MS
- MS detector setup
- Acquire
  - Match part of column that will have elution
- Segments
  - 1

Figure 1.5. Example of B- and Y- ion assignments for a peptide fragment.
Taken from Principles of Physical Biochemistry, van Holde.

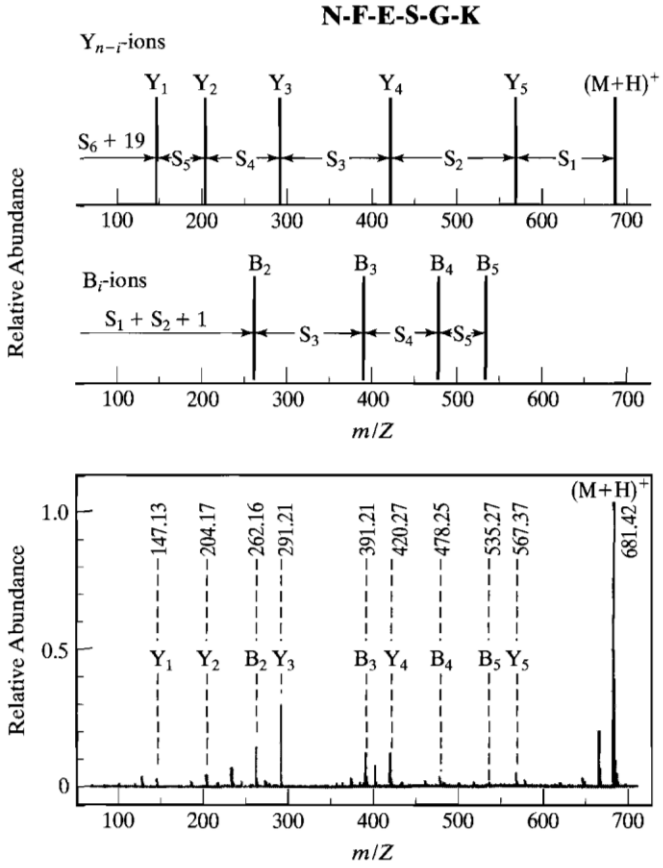


Figure 15.19 MS/MS spectrum of the peptide NFESGK. The Y- and B-ion spectra can be predicted from the molecular weights of the amino acid residue masses (Table 15.1) and the daughter ion compositions (Table 15.2). These are compared to the experimental spectrum, which includes the Y- and B-ion peaks and peaks from other daughter ions.

Tune method: angiotensin.ltqtune

C:\xcalibur\methods*.ltqtune

Scan events

1) MS; 2) MS/MS

Divert valve

3 parts: 1) waste for 2 minutes (buffer head flow); 2) sample; 3) waste (column wash)

If you haven't changed anything, do NOT save. If you changed anything, save with name + your initials; close window.

Sequence setup

File: New: Template

Base file name (will increment)

Save

Browse D:\LTQ\Users\Bell

Method: C:\xcalibur\methods\

Process methods: none

sample

inject/sample

Initial vial position: RA# (sample in red compartment, autosampler, position)
Enter
Populate list
Must set injection volume = 2 microliters (default is 25 microliters)

Run sequence icon (piece of paper with arrow from left to right)
Answers questions (no Qs really to answer), uncheck reuse vial positions
Enter
Go to queue (Left side, see that LC and MS go to "Ready to run" status)
Left side: "Acquisition Queue" tab, check all sequence, check on method to run
Click play button (Your run will take ~55 minutes)

3rd icon from left – acquisition in progress/real time plot view
Gives MS spectra, pump pressure and total ion current (chromatograph of column)
DO NOT manipulate data in this window.

1st button on left: Road map
- Choose "Qual Browser"
- Open >File
- Unpinned = static
- Pinned = dynamic
- On chromatograph, rt. Click > range
  - Choose base peak (total ion current – default)
  - Auto process – enables smoothing

SHUT DOWN
On computer, go to Chromeleon "Home" tab
Take control
Go to "HPG3400RS" tab
100% C to wash column
Run 100% C for 10 minutes.
When done, change flow to 0.0 ml/ml, disconnect; Home; Release control; Close
LTQ tune +
- Setup
- ESI source
- Capillary T = 150°C
- Apply
- Check that T begins to decrease from ~250 to 150
- Close

Turn off N_2 tank

GLOVES!!!!!
Clean MS port (Observe instructor – do NOT attempt!)
- Attach tubing to MS port.
- Open MS chamber, place Kimwipe under port to catch MeOH
- Inject 100 ml of MeOH using Hamilton syringe via connected tubing

Remove Kimwipe
Insert needle into nozzle
Close head
Detach tubing at MS port, reattach tubing from valve to MS port.

On LC

Turn all buttons to Standby mode.

***Data analyses*:**

Each member of the group will hand annotate two peptide fragments for their corresponding B- and Y- ions.

To generate a list of peptide fragments for porcine mMDH, you will need to obtain the FASTA sequence:

http://www.expasy.org/
Query database = UniProtKB
Porcine MDH
Remember: You want the sequence for mitochondrial MDH
Scroll to "Sequence" >Click FASTA >Copy to a file.

http://prospector.ucsf.edu/prospector/mshome.htm
MS-digest
Copy porcine mMDH sequence into "User Protein Sequence"
Parameters:

Digest = Trypsin; Missed cleavages = 1; Constant Mods = carbamidomethyl; Peptide mass = 350 – 4000; Min peptide length = 5; Instrument = ESI-ION-TRAP-low res
Perform Digest

Copy this list. Use this list to choose the peptide that you will analyze for B- and Y- ions.

On computer, go to Xcalibur
Click on left icon

Choose "Qual Browser"
>File>Open>***.raw (your MS data file)
The top spectrum is the total ion current (equivalent to a chromatograph of the C12 column showing elution peaks)
The bottom spectrum is the MS analysis of a selected peak in the total ion current spectrum.
On the right hand side, you can "pin" the spectra. If the spectrum is "pinned," then it is dynamic, "unpinned" is static.
Pin the top spectrum, right click >Ranges >Plot type = Base Peak >Autoprocess tab >enable smoothing checked, click OK (This applies an algorithm to the data making interpretation easier)
Pin the bottom spectrum, now you can click on any peak in the top spectrum and its corresponding MS spectrum will appear in the bottom window.
On keyboard, use right arrow to step through top window peaks – watch the bottom spectra. The first click is the MS spectrum, the 2nd click is the MS/MS of that spectrum based upon criteria set in the methods file (most intense peak or from a designated list).
For a specific peptide mass – such as those that you will hand annotate:

In the MS spectrum >right click >Ranges >Filter> find the mass in the list >OK
This takes you to the MS2 of peptide for that given mass.
To save this spectrum for analyses:

On keyboard, press "prt sc"

Open EXCEL, paste
In MS spectrum >right click >export >clipboard >Nom. Mass
Paste list into EXCEL
You now have the MS/MS spectrum for a parent peptide & a list of fragments.
You now have enough information to begin to assign the B- and Y- ions for this fragment.

EXPERIMENT 2

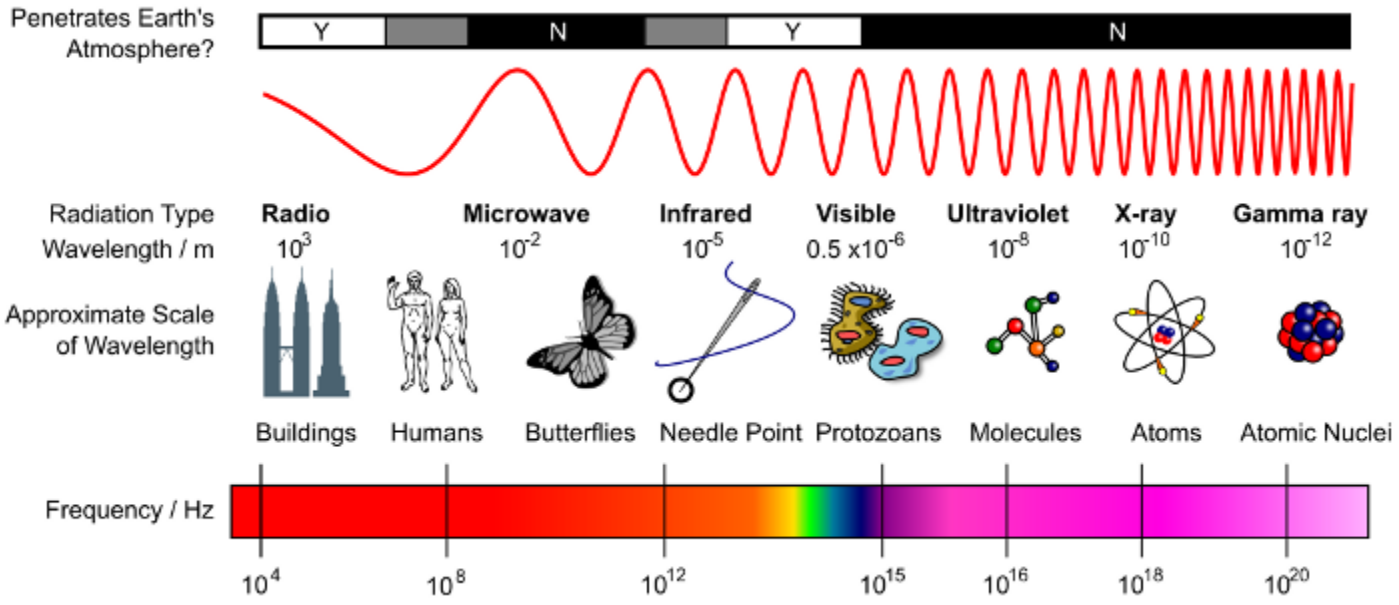
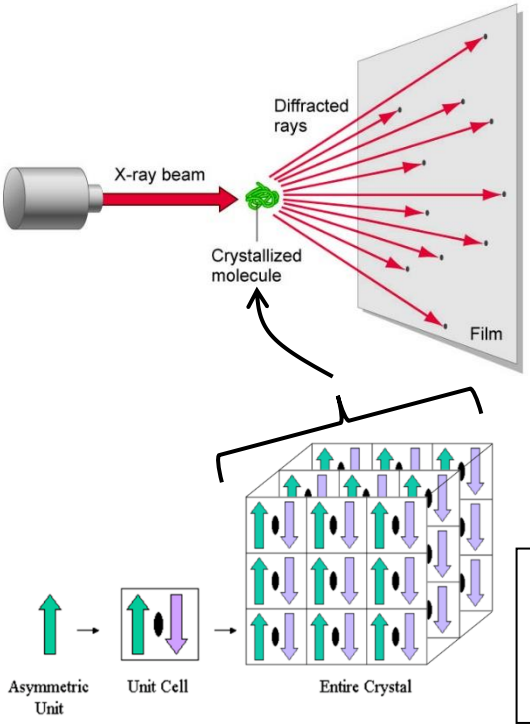
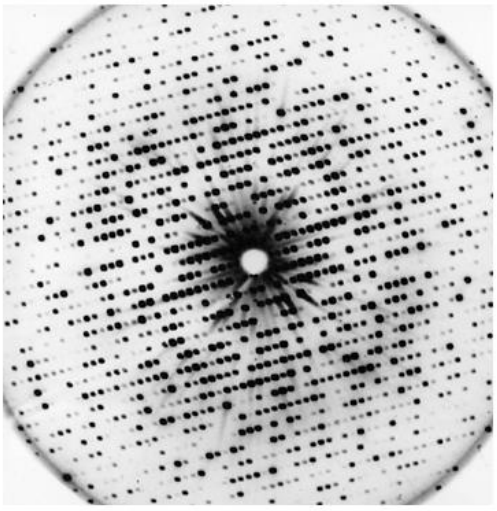


Figure 2.1 Electromagnetic spectrum.

Crystallization, crystal mounting, collection of x-ray diffraction data and data analyses for lysozyme.

To resolve an object, you must use the range of the electromagnetic spectrum that produces wavelengths of the same magnitude as the size of object to be resolved. For proteins, formed primarily from bonding patterns of C, N, O and H, that are on the order of 1.3-2.5 angstroms (0.13-0.25 nm), this corresponds to the wavelength of x-rays, see Figure 2.1. Unlike visible light, x-rays have a refractive index less than 1. Thus, instead of using transparent material to focus x-rays, the process of reflection, diffraction or interference is used to measure the interaction of x-rays with an object. In structural biology, diffraction of x-rays from the electrons of a protein provides information about the protein's structure. Diffraction from a single protein molecule is very weak, unmeasurable. Instead, multiple copies of the protein held in a fixed array known as a crystal act as the sample. The periodic spacing of the protein molecules produces, at defined orientations of the crystal to the x-ray, constructive (reflected x-rays are in phase with one another) or destructive (x-rays are out of phase with one another) diffraction. The pattern of

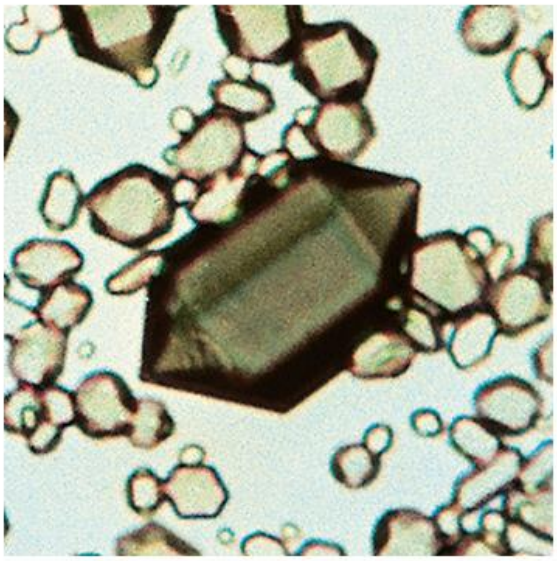




Courtesy of John Kendrew, Cambridge University, U.K.

Figure 2.2. Overview of x-ray diffraction. A protein crystal is bathed in an x-ray beam. The x-rays are diffracted off of the protein's electrons. The diffraction pattern is captured by the detector.

diffracted x-rays is captured on a detector, recording the intensity (amplitude) of the x-ray wave. This diffraction pattern describes the electron structure of the crystal lattice. To interpret this data, one must know the wavelength, amplitude and phase of the x-rays that formed the diffraction pattern. Crystallographers define the wavelength of the experiment by the method of generating x-rays. For example, electrons are generated and released at high velocity. When they collide with an electron rich surface, such as pure copper, the collision provides enough energy to eject an electron from an inner shell. When a higher orbital electron fills this inner shell vacancy, energy is released in the form of x-rays at a defined spectral lines (i.e. wavelengths) that are dependent upon the type of electron rich surface. The amplitude is directly quantitated by the intensity of the diffracted x-rays. Only the phase cannot be directly measured but must be determined by direct methods, multi- or single wavelength anomalous diffraction, single or multiple isomorphous replacement, or molecular replacement.



Courtesy of John Olsen, Brookhaven National Laboratories, and Brian Matthews, University of Oregon.

Figure 2.3. Crystals in mother liquor.

Introduction to Crystal growing

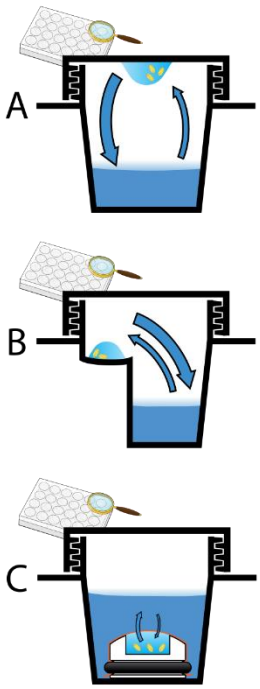


Figure 2.4. Approaches to crystallization. A) Hanging drop vapor diffusion. B) Sitting drop vapor diffusion. C) Microbatch. "Crystal Drops" by Adenosine - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons - http://commons.wikimedia.org/wiki/File:CrystalDrops.svg#mediaviewer/File:CrystalDrops.svg

The art of crystal growing is almost a black magic, requiring not only extremely pure reagents and excellent laboratory technique, but a good deal of luck to obtain crystals suitable for x ray crystallography. In the current laboratory, you will attempt to obtain crystals of lysozyme.

Proteins are crystallized from a supersaturated solution. To achieve super saturation, proteins are incubated with precipitants that effectively sequester water thereby reducing the protein's interaction with solvent and reducing its solubility/increasing its effective concentration. Common precipitants are salts, like ammonium sulfate, or polyethylene glycols of defined molecular weight. In addition to precipitants, solubility of proteins can be altered by pH and temperature. At a pH near its isoelectric point, the near neutral protein makes fewer interactions with its polar solvent and more readily precipitates from solution. Temperature can be used to increase the rate of precipitation (higher T) or slow the rate precipitation (lower T).

To identify a solution compatible with protein crystallization, one completes an incomplete factorial screen. An incomplete factorial screen is a carefully chosen subset of all combinations of buffers, salts, precipitating agents available. The subset is chosen so as to exploit common properties observed to result in crystals, but limit the number of conditions sampled – thereby decreasing the amount of sample needed for the experiment. A typical screen will contain 96 solutions that have the following properties, pH 4-9, 0.1-2 M salts, 10-40% polyethylene glycols (2 distinct molecular weights), 0.1 M common cations and anions (i.e. Mg^{2+}, Zn^{+}, Ca^{2+}, $SO4^{-2}$, Cl^{-}), up to 10% organic solvents (i.e. 2-methyl-2,4-pentanediol).

In the growth of lysozyme crystals, the optimal conditions involve pH 4.7

and a buffer containing 0.83M NaCl in 0.2M Acetate. The conditions will vary according to the preparation of the lysozyme, the concentration of the protein and the temperature. Ideal crystal growing conditions involve a limited number of nucleation centers and slow crystal growth so that relatively large crystals can be obtained.

To grow crystals, two approaches are commonly taken: vapor diffusion or microbatch, see Figure 2.3. In both vapor diffusion approaches, the solution containing the buffer, salt, precipitant or other additives, called the "well" solution, is placed in the bottom of the crystallization plate, termed the well. The protein, typically concentrated to ~10 mg/ml in a weak buffer (5-10 mM), is mixed 1:1 with the well solution and placed on a coverslip (in hanging drop) or on the shelf (sitting drop). The well is then sealed and the drop is allowed to equilibrate with the well solution. Because the protein drop is ½ as concentrated as the well solution. The precipitants in the well solution will absorb water from the protein drop, thereby concentrating the protein solution over time. In the microbatch setup, the solution in the well is typically oil like paratone. To the base of this solution, 1:1 solution of protein and well solution is pipetted. Over time water diffuses/evaporates from the oil layer. The rate of water loss is controlled by the oil composition. In this laboratory, you will use two different approaches to attempt to obtain crystals of lysozyme, a batch approach and the vapor diffusion approach.

Growing Lysozyme Crystals by Microbatch

A series of salt and protein concentrations are screened in the batch protocol, which is simple to set up, but relatively costly in terms of the amounts of protein that are required. Two solutions must first be prepared:

Solution I
0.1M NaCl
0.2M Acetate Buffer, pH 4.7
40mg/mL Lysozyme

The solution is filtered through a 0.2μ Cellulose Acetate Syringe Filter to remove any particulate material.

Solution II

2M NaCl
0.2M Acetate Buffer, pH 4.7
5mM β–Mercaptoethanol

The solution is filtered through a 0.2μ Cellulose Acetate Syringe Filter to remove any particulate material.

Protocol:

To the microbatch plate, add 10 ml paratone. Make sure that all wells are filled with oil. Remove any air pockets. Prepare 5 wells of the same condition. Pipette solutions as indicated in Table 2.1.

Table 2.1. Microbatch Set Up.

Vial #	*Solution I*	*Solution II*	*Protein*	*[NaCl]*
#	μl	μl	mg/ml	M
1	10	5		
2	10	7		
3	10	8		
4	10	9		
5	10	10		

The microbatch plate is store at room temperature for a minimum of 2 days.

Growing Lysozyme Crystals by Vapor Diffusion

3 Solutions are required:
1M Acetate
1M NaOH
5M NaCl

These solutions are mixed in varying amounts and brought up to 1mL with H_2O to give a variety of conditions varying pH and salt concentration according to the accompanying table.

To achieve a scan of pH and NaCl concentrations use the following amounts, made up to a total volume of 1mL with water.

Table 2.2. pH and Salt Concentrations.

pH of Well	µL 1M Acetate	µL 1M NaOH	NaCl Concentration	µL 5M NaCl
3.6	200	20	0.1M	20
3.7	200	22	0.2M	40
3.8	200	24	0.3M	60
3.9	200	30	0.4M	80
4.0	200	36	0.5M	100
4.1	200	44.4	0.6M	120
4.2	200	52.8	0.7M	140
4.3	200	63.6	0.8M	160
4.4	200	74	0.9M	180
4.5	200	86	1.0M	200
4.6	200	98	1.2M	240
4.7	200	110	1.4M	280
4.8	200	118	1.6M	320
4.9	200	129	1.8M	360
5.0	200	140	2.0M	400
5.1	200	149		
5.2	200	158		
5.3	200	165		
5.4	200	172		
5.5	200	177		
5.6	200	182		

From the above table you should select **4 pH values** and **6 NaCl concentrations** and set up a 24 well grid. The appropriate solutions are then added to each well of a 24 well microplate, and made up to 1mL with the appropriate

volume of H_2O.

Mix the solutions in each well until they are uniform: ie no visible phase separation.

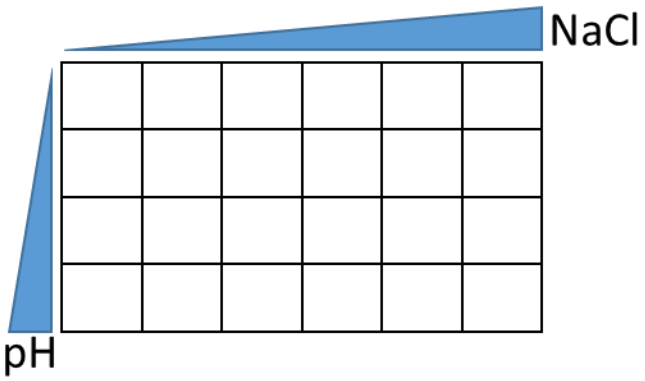
Pipet 3μL of the well solution to the center of a siliconized, dust-free, lint-free cover slip. Keep the drop as round as possible and bubble free.

Add 3μL of the protein solution to the drop: **DO NOT MIX.** You will create bubbles if you try to mix. Mixing by diffusion occurs rapidly in such a small volume.

Invert and press the cover slip down on the greased well. Be careful not to break the cover slip, but make sure that you have a good seal.

Incubate the plate undisturbed for 4-5 days at room temperature.

During the second lab period, we will mount & observe the x-ray diffraction of your crystals.



EXPERIMENT 3

Computation approaches to study macromolecular structure

It is a central tenant of molecular biology dogma that the nucleotide sequence of a gene governs the amino acid sequence of the coded protein and that the amino acid sequence of a protein governs the structure, and hence function of the resultant protein. We will explore the bioinformatics tools that allow information to be deduced from the amino acid sequence of a protein. In turn, we will address the Prediction of Physical Properties, the Identification of Potential Function, the Prediction of Secondary and Tertiary Structure, the Predication of Flexibility and Dynamic Properties of a protein, and the Identification of Potential Sites of Post-Translational Modification. Since a critical aspect of such bioinformatics approaches involves "homology," we start with a brief discussion of homology in protein sequences, and conclude with a section of how analysis of homology (in its broadest sense involving either multiple sequences or multiple structures) using bioinformatics tools can give further insight into protein structure function relationships.

What is homology?

Sequence identity is exact homology. Closely related proteins will have a high degree of sequence identity. More distantly related proteins will have a lesser degree of sequence identity, but may also show sequence conservation where residues of similar types [physical properties and function] have been substituted. More distantly related proteins will also often have gaps or inserts where regions of sequence have been deleted or added. The algorithms that calculate homology usually have some type of penalty for gaps in the alignment since it is, of course, easier to align two or more sequences if you allow gaps of unlimited size.

Similar sequences can be identified via BLAST searches (http://blast.ncbi.nlm.nih.gov/Blast.cgi) or OrthoDB (http://orthodb.org/).

DO:

Go OrthoDB and identify 20 sequences for the protein, Suppressor of IKKepsilon (SIKE).
For OrthoDB,
Select the "Build your query" tab:
Text = SIKE
Phyloprofile: no filtering; single copy >90% species
Select species: select bacteria, fungi & metazoan
Submit

Choose Group with the most genes in the most species
Click "View FASTA"
Copy into a file.
Note: In the FASTA sequence list, you will have proteins in this format:

```
>69293:00118c ENSGACP00000006028 gene=ENSGACG00000004553 orthodb8_OG=EOG85QJC3
orthodb8_level=1261581 organism_name=`Gasterosteus aculeatus` uniprot_de=`Uncharacterized
protein`
MACTMDKVLGDARTLLERLKEHDLAAEGLIEQSGALSQRVQGMKEVGNALPDKHTEETSEIQELLKFKPHVLLAQENTQIKDLQQENKE
LWLSLEEHQYALELIMGRYRKQMLQLMMAKKELDTKPVLSLHENHAKEVQSQVERICEMGQVMRRAVQVDDQHYCSVAERLAQLEIENK
ELRDLLAISKSSVKAAREESSQPTAAPPPQPGPHE
```

You may want to make an additional copy of the sequences and rename to use the species name as the identifier, as follows:

```
>G_aculeatus
MACTMDKVLGDARTLLERLKEHDLAAEGLIEQSGALSQRVQGMKEVGNALPDKHTEETSEIQELLKFKPHVLLAQENTQIKDLQQENKE
LWLSLEEHQYALELIMGRYRKQMLQLMMAKKELDTKPVLSLHENHAKEVQSQVERICEMGQVMRRAVQVDDQHYCSVAERLAQLEIENK
ELRDLLAISKSSVKAAREESSQPTAAPPPQPGPHE
```

Sequences that have long stretches of X (unidentified amino acids) should be excluded from your list of protein sequences, such as:

```
>37347:001a49 ENSTBEP00000006722 gene=ENSTBEG00000007772 orthodb8_OG=EOG85QJC3
orthodb8_level=1261581 organism_name=`Tupaia belangeri` uniprot_de=`None`
MSCTIEKILTDAKTLLERLREHDAAAESLVDQSAALHRRVAAMREAGTALPDQVRQRYQEDASDMKDMSKYKPHILLSQENTQIRDLQQ
ENRELWVSLEEHQDALELIMSKYRKQMLQLMVAKKAVDAEPVLKAHQSHSAXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
LENKELRELLSISSESLQARKENTMDTVSQAIK
```

These sequences (in FASTA format - http://en.wikipedia.org/wiki/FASTA_format) can be aligned using a program such as Clustal Ω. In these types of programs, it is usually possible to color code residue types and to mark exact homology and functional homology. The standard four colors used to highlight amino acid residues are shown in Table 3.1. While Table 3.1 showed electrostatic and polarity issues, physical size is also an important property to consider, see Table 3.2.

Table 3.1. Standard formatting for CLUSTAL Ω sequence alignments.

Color	Amino Acid Side Chains (pK)	Conserved Property	Comments
Purple	**K(10.5), R(12.5)**	**Positive charge**	
Blue	**E(4.1), D(3.9)**	**Negative Charge**	Q and N are neutral analogs
Green	**G,S,T,Q,N,H(6.0),Y(10.5),C(8.4)**	**Hydrophilic**	H can be +ve, Y or C can be -ve
Red	**M,L,I,V,A,F,W,P**	**Hydrophobic**	P restricted Φ angle

Table 3.2. Physical properties of amino acids considered in homology.

Group	Sidechain	Volume, A^3		Group	Sidechain	Volume, A^3
Small	Glycine	66		**Medium-Small**	Threonine	122
	Alanine	92			Aspartate	125
	Serine	99			Proline	129
	Cysteine	106			Asparagine	135
					Valine	142
Large	Tyrosine	203		**Medium-Large**	Glutamate	155
	Phenylalanine	203			Glutamine	161
	Arginine	225			Histidine	167
	Tryptophan	240			Leucine	168
					Isoleucine	169
					Lysine	171
					Methionine	171

DO:

Go to CLUSTAL Ω (http://www.ebi.ac.uk/Tools/msa/clustalo/)

Paste your SIKE sequences into the input sequences box.
Submit
Examine the sequence alignment. Select "Show colors" to quickly identify regions of similarity.
Copy/Download your alignment file.
Examine the Phylogenetic tree.

DO:

How far apart are your sequences in evolutionary terms?

Use of Time Tree (http://www.timetree.org/index.php) web project to identify the largest evolutionary distance within your sequences.

By comparing a sequence with known "consensus sequences," it is possible to identify an active site motif, binding domain, or potential sites of post translational modification. For example running "Prosite" [http://us.expasy.org/prosite/] with the glyoxosomal MDH sequence will give an indication of "possible" sites of N (Asn)-linked glycosylation [NxS/T], Protein Kinase C phosphorylation [TxR], Casein Kinase 2 phosphorylation sites [S/TxxD/E], Tyrosine Phosphorylation [RxxxExx.Y], myristoylation [GGxxGF], RGD cell attachment sequences [RGD] and of course the malate dehydrogenase active site signature sequence [VTMLDxxRAxxxV]. You should note that, apart from the malate dehydrogenase active site signature sequence, there is no evidence of any of the other predicted sites being experimentally verified sites: they are simply sites that have the appropriate consensus sequence.

A variety of "prediction" tools available through ExPASy:

http://www.expasy.org/proteomics

DO:

Go to the proteomics site of ExPASy
Choose 10 programs from the list of tools and examine SIKE using these programs.

What is Bioinformatics?

Bioinformatics, often referred to as computational molecular biology, is an interdisciplinary field. Bioinformatics develops methods and software tools for understanding biological data in its broadest sense. It uses the data to predict properties of molecules or biological systems and explores relationships in the large data sets of biological information that are now available as a result of the various genome projects. Bioinformatics combines computer science, statistics, mathematics, and engineering principles to explore biological data.

The utilization of bioinformatics tools is facilitated by the existence of several repositories of web-based, peer reviewed, validated tools including the ExPASy suite maintained by the Swiss Bioinformatics Institute, (www.expasy.org), the European Molecular Biology Laboratory (EMBL) European Bioinformatics Institute (EBI), (http://www.ebi.ac.uk/services), and the University College London Bioinformatics group (http://bioinf.cs.ucl.ac.uk/web_servers/).

Together with web-based, open access repositories of biological data, these tools are freely accessible to the research and educational communities of the world. Two sites, in particular, are useful for the application of bioinformatics approaches to protein structure and function: the National Center for Biotechnology Information (NCBI) accessible through PubMed, (http://www.ncbi.nlm.nih.gov/pubmed) where protein (and nucleotide) sequences can be readily obtained and the Protein Data Base (PDB), which is the repository of three dimensional coordinates for macromolecules determined by X Ray Crystallography or NMR approaches (http://www.rcsb.org/pdb/home/home.do)

Prediction of Physical Properties

DO:

Using the ExPASy site:
http://www.expasy.org/proteomics
Submit SIKE to three different programs that provide physical properties. Hint: ProtParam is a good start.

Identification of Potential Function

DO:

1) Run PROPSEARCH on the SIKE sequence

   What function would you predict based upon the output?

2) Examine the entry for human SIKE in the UniPROT database.

   What information can derive about SIKE function from this entry?

Prediction of Secondary and Tertiary Structure

Secondary structure may be defined as the local spatial organization of the polypeptide backbone without consideration of the side-chain conformations. As we will see, however, when considering the prediction of secondary structure from the amino acid sequence of the protein, the nature of the side chains in a particular region of polypeptide chain does influence whether a certain secondary structure is found. The secondary structure is defined by four basic categories: α helix, β strand (often associated into so-called "sheets", oriented in a parallel or an anti-parallel manner), β turn, and random coil.

The *tertiary structure* of a protein is defined as the packing of the foregoing secondary structural elements *within* a polypeptide chain into a three-dimensional structure. Although as just defined, a tertiary structural element should involve a -single polypeptide chain, there are instances where an apparent tertiary structural element involves two or more polypeptide chains.

Secondary Structure Prediction

Because the prediction of secondary structure is based largely on the character of the amino acids that are found in nature to be in certain of the secondary structures, it is informative to consider briefly how procedures for these predictions from primary sequence have been developed and applied. Many of the problems are similar to those the protein itself must encounter during the folding process! Predictive methods are based on the probability that a particular type of amino acid residue is found in a certain type of secondary structure. These data are obtained in one of two ways:

Table. 3.3. Amino acid preferences in forming secondary structures.

Amino acid	α-helix (Pα)	β-sheet (Pβ)	Turn (Pt)	
Ala	1.29	0.90	0.78	Favor α-helices
Cys	1.11	0.74	0.80	
Leu	1.30	1.02	0.59	
Met	1.47	0.97	0.39	
Glu	1.44	0.75	1.00	
Gln	1.27	0.80	0.97	
His	1.22	1.08	0.69	
Lys	1.23	0.77	0.96	
Val	0.91	1.49	0.47	Favor β-sheets
Ile	0.97	1.45	0.51	
Phe	1.07	1.32	0.58	
Tyr	0.72	1.25	1.05	
Trp	0.99	1.14	0.75	
Thr	0.82	1.21	1.03	
Gly	0.56	0.92	1.64	Favor turns
Ser	0.82	0.95	1.33	
Asp	1.04	0.72	1.41	
Asn	0.90	0.76	1.23	
Pro	0.52	0.64	1.91	
Arg	0.96	0.99	0.88	

In the first, probabilities are obtained by examining the crystal structures of known proteins and counting the number of times particular residues appear in *a* helices, *β* strands, or *β* turns. Alternatively, polymers of single amino acids are used and their secondary structure determined. From the tendency of such polymers and various copolymers to form *a* helix and *β* sheet, an assessment of the contribution of individual residues to these structures can be made. In some instances it is useful to keep information on *where* in the type of secondary structure these residues appear most frequently, as this can be helpful in defining starting points and termination points for the type of secondary structure. Table 3.3 gives such information concerning probabilities of residues appearing in *a* helices, *β* strands, and *β* turns.

Several generalities can be drawn from Table 3.3: (1) the charged residues are unfavorable for β-strand formation, and three of them (Asp, His, Arg) are also α-helix indifferent; (2) residues that tend to break α helices (Pro, Gly, Asn, Tyr) also tend to be residues with high probability of appearing in β turns, and (3) residues with a strong tendency to be in β strands are rarely found in β turns. This type of information has been applied to secondary-structure prediction. In a predictive scheme the influence on neighboring residues is taken into account in attempting to assign a propensity of each residue in a peptide to be in an α helix, a β strand, or a β turn. Each type of secondary structure is "predicted" independently and the final "prediction" based on a comparison not only of the "strength" of the prediction, but also on the predictions for adjacent residues. For example, it is quite possible for a region of peptide to contain a residue that has a high probability of being in either a β strand or an α helix; if the neighboring residues are predominantly helical, this weights the final choice between β strand and α helix for the prediction. Finally, regions of secondary structure are predicted based on certain "nucleation" rules. For an α helix to be indicated six adjacent helical residues must be present, for a β strand to be indicated five adjacent strand residues must be present, and for a β turn two residues, of a tetrapeptide sequence, must be indicated as strong turn formers. With β turn predictions, weighting is given to proline in the *second* position in the turn.

DO:

Submit the SIKE sequence to the following secondary structure prediction programs:

http://bioinf.cs.ucl.ac.uk/psipred/
http://www.compbio.dundee.ac.uk/www-jpred/

http://embnet.vital-it.ch/software/COILS_form.html
http://dis.embl.de/

Compare the results between the programs. Do they agree?

Tertiary Structure Prediction

The explosion of genome sequencing projects in recent years means that the percentage of known proteins (at the sequence level) that have corresponding three dimensional structures experimentally determined has continued to shrink to less than 1% currently. With increasing emphasis at interpreting data at the molecular and atomic levels of detail, experimental determination of three dimensional structures cannot keep up. Combined with the fact that many interesting proteins (for example membrane proteins and proteins containing intrinsically disordered regions) are very difficult to crystallize, it is likely that prediction of three dimensional structures will be increasingly important as a research tool. Two basic paradigms for the prediction of tertiary structures are currently used, one based upon evolutionary relationships between proteins and the other based upon ab initio calculations using established principles governing protein folding. Resultant models must be validated in a variety of ways and the "accuracy" of such models limits their usefulness and can be categorized into three basic groups. High resolution models can be used for detailed mechanistic interpretation of protein structure-function relationships while medium resolution models can help define active site regions, ligand binding sites and potential roles of mutations. Low level models may give insight into overall topology and can be used to tentatively define protein boundaries or assign a protein to a particular family or superfamily. Three dimensional predictions can also be used to help interpret experimental data and even solve experimental three dimensional structures using x ray data (diffraction or scattering), electron microscopy, or NMR.

Tertiary structure prediction approaches are based on several foundational concepts developed at the start of the molecular era of protein structure. Before going into detail of how predictions are currently done, it is important to put these approaches into a historical context. On the basis of the first crystal structure of myoglobin, Kendrew formulated four generalizations that still guide thinking about tertiary structure:

1. Proteins are compact structures having very small amounts of internal solvent molecules, which are present internally and presumed to have been trapped during the folding process.

2. Almost all the polar side chains in the protein are at the surface of the molecule, where they can interact with solvent and solute molecules in the bulk solvent. Any exception to this would indicate that a "non-surface" polar group is involved in some internal function. In the case of myoglobin, for example, a histidine side chain is buried internally, but is associated with the heme ring of the molecule.

3. All nonpolar residues, with the possible exceptions of glycine and alanine, are located in the interior of the molecule. Glycine and alanine, because of their "short" side chains, can be located at the surface.

4. All polar groups at the surface of the molecule, whether they are side-chain or main-chain C=O and N-H groups, have bound water molecules.

The relationship between amino acid sequence, tertiary structure, the idea of a protein being at a global free energy minimum, and the time scales of molecular motions involved in folding led to "Levinthal's Paradox" and the

suggestion that the folding process itself cannot be completely random since it would take too long relative to the overall life of the protein. A relatively small protein of, for example, 100 amino acids has approximately 10^{40} possible conformations, depending on its primary structure. Small proteins generally fold on a time scale of seconds. The molecular motions involved in the folding of a polypeptide chain occur on a nanosecond-to-picosecond time scale, suggesting that at most about 10^{11} conformations could be randomly screened during the folding process. These considerations led Levinthal to suggest kinetic control of folding and subsequent ideas about folding funnels.

The relationship between sequence and structure (albeit secondary structure) led to the Chou and Fasman concepts of secondary structure prediction which is still an integral step in tertiary structure prediction approaches. Finally, it is important to recognize that many of the approaches discussed here are based in some way on the evolutionary relationships between proteins and the concepts of functional domains in proteins. The energy of a particular conformational state of a protein can be computed and in some way minimized to an energy minimum where both short and long range interactions are considered to give a structure resembling the native state of the protein.

Homology Modeling.
Over the years a variety of programs that predict tertiary structures based upon evolutionary relationships at the level of sequence have been developed including programs such as "Modeller" (https://salilab.org/modeller/), and "Swiss-Model" (http://swissmodel.expasy.org/).

Other prediction methods include: MULTICOM, I-TASSER, and Robetta/Rosetta. The approaches or output of these programs are summarized in Figure 3.1-3.3. Figures 3.4-3.5 outline programs that redesign existing protein structure for an alternative function or predict function based on homology to other 3-D macromolecular structures.

DO:
Model the structure of SIKE.
Use the programs in the figures or the ExPASy website:
http://www.expasy.org/structural_bioinformatics
Click on Protein Model Portal. Submit a modelling request to at least two web programs.
Use PYMOL to evaluate your models. Do they agree? Submit your model files to the class DROPBOX.

Figure 3.1. The Organization of the "MULTICOM" toolbox.
(http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html)

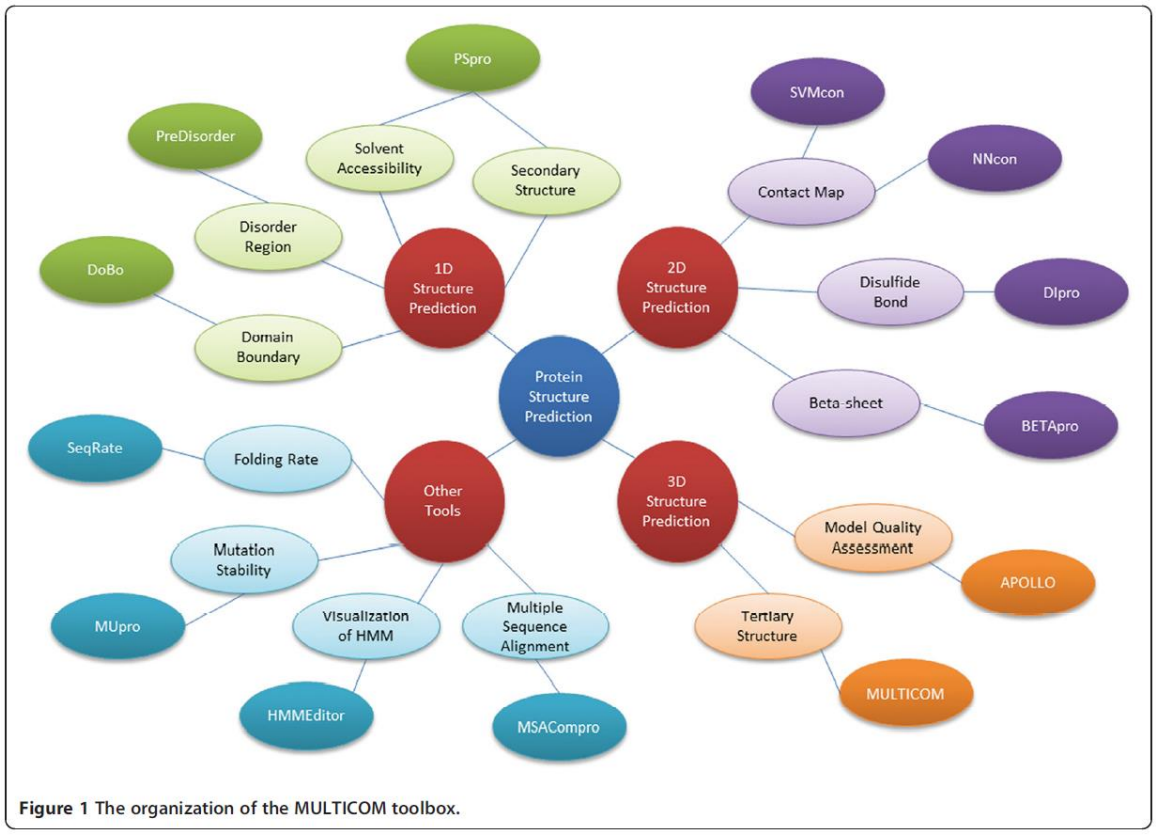


Figure 1 The organization of the MULTICOM toolbox.

Figure 3.2. I-TASSER Protocol – a threading approach. (http://zhanglab.ccmb.med.umich.edu/I-TASSER/)

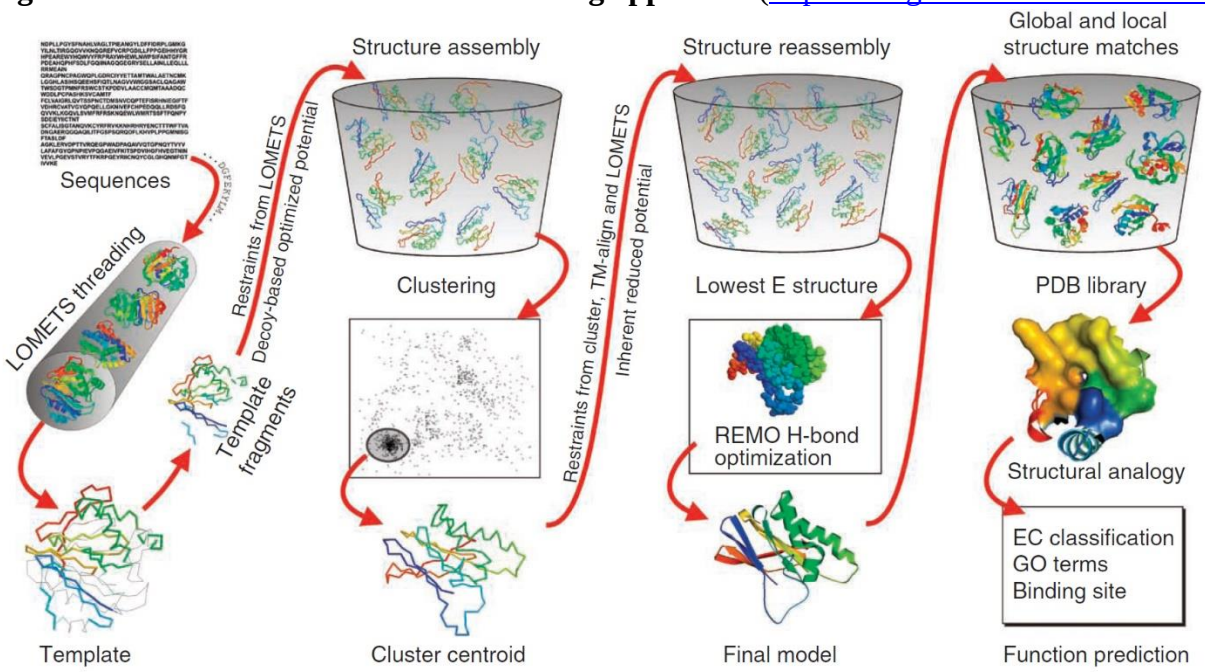


Figure 3.3 Robetta/Rosetta De Novo Folding Algorithm – an ab initio structure approach. (http://robetta.bakerlab.org/)

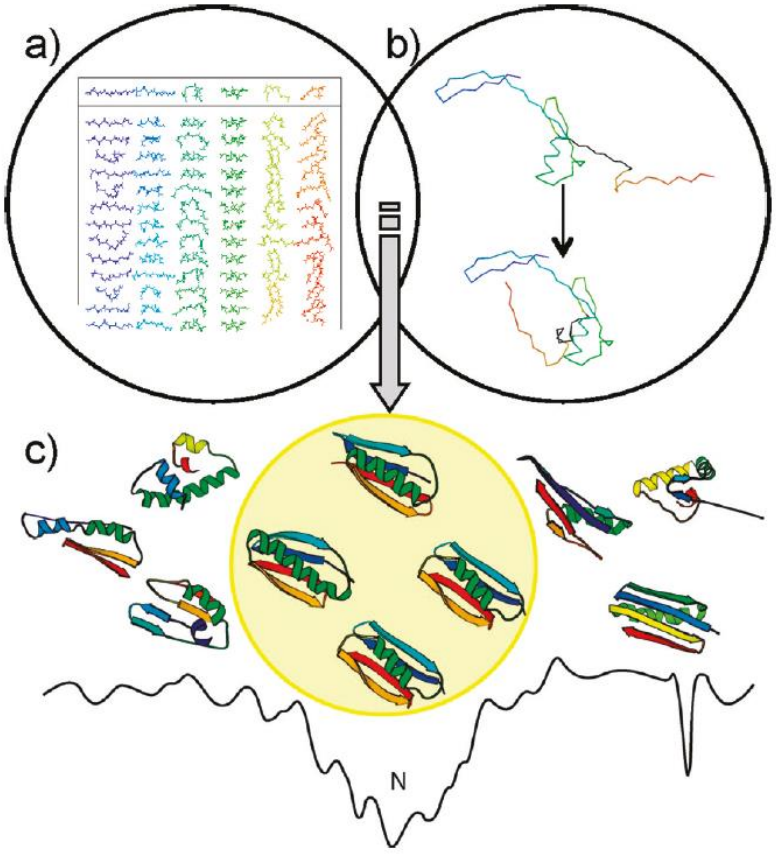


FIGURE 1: *De novo* folding algorithm. ROSETTA starts from (a) fragment libraries with sequence-dependent (φ and ψ) angles that capture the local conformational space accessible to a sequence. (b) Combining different fragments from the libraries folds the protein through optimization of non-local contacts. The low-resolution energy function depicted in panel c smoothes the rough energy surface, resulting in a deep, broad minimum for the native conformation. Metropolis Monte Carlo minimization drives the structure toward the global minimum.

Figure 3.4. RosettaDesign – Redesign a protein scaffold. (http://rosettadesign.med.unc.edu/)

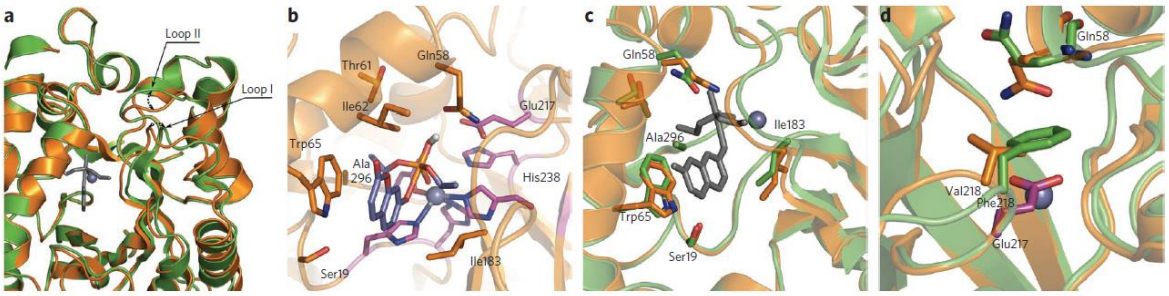


Figure 4 | Design model and crystal structure of apoPT3.1. (**a**) Superposition of the PT3.1 design model (gold) and the crystal structure (green) shows that although the overall backbone similarity is high (backbone r.m.s. deviation = 0.65 A), there are small shifts in two active site-proximal loops. (**b**) The PT3 design model showing alignment of the transition state in the redesigned enzyme. (**c**,**d**) Side chains of designed residues Ser19, Trp65, Ile183 and Ala296 observed in the crystal structure (green) adopt rotamers predicted in the designed model (gold, **c**), whereas the side chain of residue Gln58 adopts a different rotamer in the crystal structure (green) compared to the designed rotamer (gold, **d**), which corresponds to a change in its χ_2 dihedral angle.

Figure 3.5. COFACTOR – predict function from structure
(http://zhanglab.ccmb.med.umich.edu/COFACTOR/help.html)

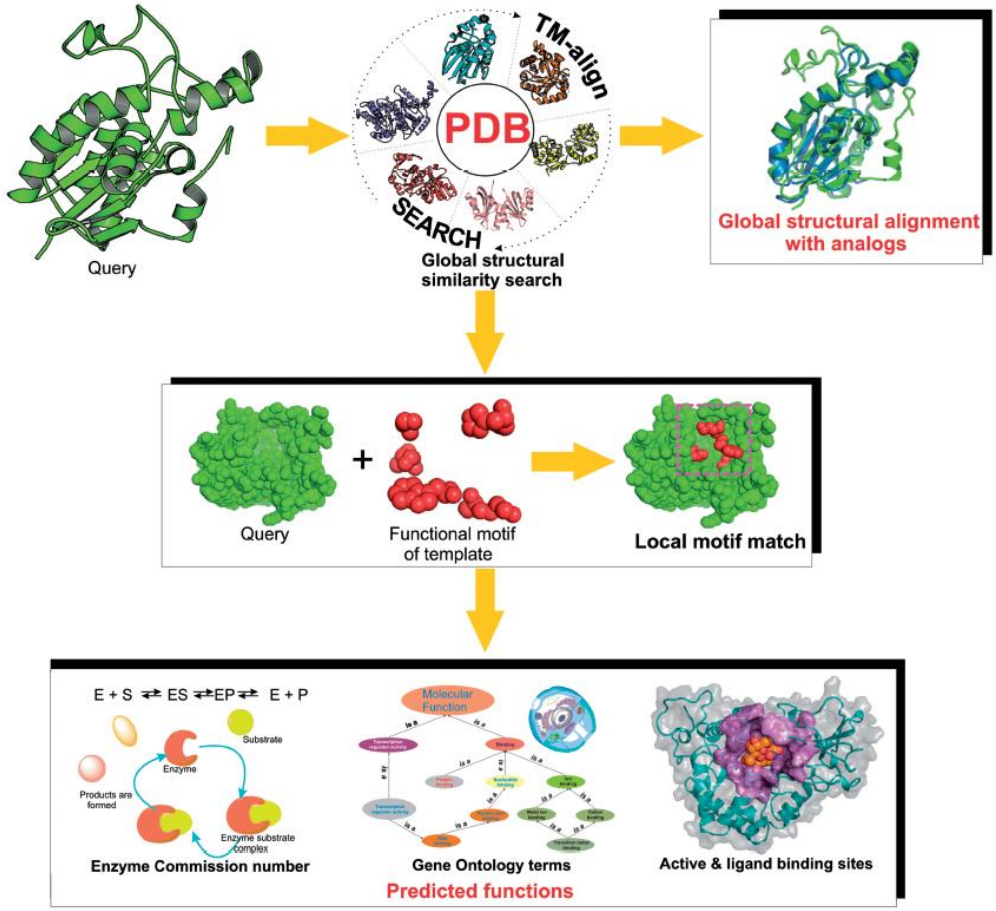


Figure 1. Illustration of structure-based function annotation by the COFACTOR server, starting from the query structure (shown in green).

References:

1. ANTONY W. BURGESS AND HAROLD A. SCHERAGA "Assessment of Some Problems Associated with Prediction of the Three Dimensional Structure of a Protein from Its Amino-Acid Sequence" Proc. Nat. Acad. Sci. USA, Vol. 72, No. 4, pp. 1221-1225, April 1975.
2. Kristian W. Kaufmann, t Gordon H. Lemmon, t Samuel L. DeL uca, t Jonathan H. Sheehan, t and Jens Meiler, "Practically Useful: What the ROSETTA Protein Modeling Suite Can Do for You" *Biochemistry* 2010, *49,* 2987-2998 2987.
3. Jianlin Cheng, Ji long Li\ Zheng Wang , Jesse Eickholt and Xin Deng "The MULTICOM toolbox for protein structure prediction" , *BMC Bioinformatics* 2012, 13:65.
4. Ambrish Roy, Alper Kucukural & Yang Zhang "1-TASSER: a unified platform for automated protein structure and function prediction: NATURE PROTOCOLS, VOL.5 N0.4, 2010, 725 -728.
5. Ambrish Roy, Jianyi Yang and Yang Zhang "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation", *Nucleic Acids Research, 2012, 1-7.*
6. Jens Meiler and David Baker "Coupled prediction of protein secondary and tertiary structure" PNAS, 2003, vol. 100, 12105-12110.
7. James Thompson, and David Baker, "Incorporation of evolutionary information into Rosetta comparative modeling" Proteins: Structure, Function, Bioinformatics (2011) 2380-2388.
8. N. MICHAEL GREEN and MICHAEL T. FLANAGAN, "The Prediction of the Conformation of Membrane Proteins from the Sequence· of Amino Acids", Biochem J., (1976) 153, 729-732.
9. Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien treutlle, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, J. Fleishman, Jacob E. Corn, David E. Kim. Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popovic, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker *and* Philip Bradley "ROSETTA3: AN OBJECT-ORIENTED SOFTWARE SUITE FOR THE SIMULATION AND DESIGN OF MACROMOLECULES" *Methods in Enzymology,* Volume 487 , 545-574.
10. Lawrence A Kelley & Michael J E Sternberg "Protein structure prediction on the Web: a case study using the Phyre server" NATURE PROTOCOLS, VOL.4, 2009, 363 -367.
11. Pedro Silva, "Assessing the reliability of sequence similarities detected through hydrophobic cluster analysis" Proteins 2008; 70:1588-1594.
12. Z. Yang, K. Lasker, D. Schneidman-Duhovny, B. Webb, C. C. Huang, E.F. Pettersen , T.D. Goddard , E.C. Meng, Andrej Sali, T. E. Ferrin, "UCSF Chimera, MODELLER, and IMP: An integrated modeling system" Journal of Structural Biology 179 (2012) 269–278.
13. J. E. Bell, E. Cunningham, C. Belt, J. D. B. Featherstone and J. Bell, "EXAMINATION OF THE POTENTIAL STRUCTURE OF HUMAN SALIVARYCYSTATINS BASED ON COMPUTER MODELLING" Arch. Oral Biol. Vol. 42,. 761 772.
14. Konstantin Arnold, Lorenza Bordoli, Jurgen Kopp, and Torsten Schwede,"The SWISS-MODEL workspace: a web-based environment for protein structure homology modeling" BIOINFORMATICS, Vol. 22, 2006, pages 195–201.
15. Levinthal C. 1969." How to fold graciously". In: Debrunner P, Tsibris JCM, Munck E, eds. *Mossbauer spectroscopy in biological systems* ~Proceedings of a meeting held at Allerton House, Monticello, Illinois!. Urbana, Illinois: University of Illinois Press. p 22.
16. K. Lasker, J.L. Phillips, D. Russel, J. Velazquez-Muriel,D. Schneidman-Duhovny, E. Tjioe, B. Webb, A. Schlessinger, and Andrej Sali, "Integrative Structure Modeling of Macromolecular Assemblies from Proteomics Data" *Molecular & Cellular Proteomics 9.8* 1689-1702.
17. Mark E. Peterson, Feng Chen, Jeffery G. Saven, David S. Roos, Patricia C. Babbitt, and Andrej Sali "Evolutionary constraints on structural similarity in orthologs and paralogs" PROTEIN SCIENCE 2009 VOL 18:1306—1315.
18. Sagar DKhare, Yakov Kipnis, Per Jr Greisen, Ryo Takeuchi, Yacov Ashani, Moshe Goldsmith, Yifan Song, Jasmine LGallaher, Israel Silman, Haim Leader, Joel LSussman, Barry LStoddard, Dan S Tawfik & David Baker,

"Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis", Nature Chemical Biology, vol 8, 2012, 294-300.

19. Ambrish Roy, Jianyi Yang and Yang Zhang, "COFACTOR: an accurate comparative algorithm for structure-based protein function annotation" Nucleic Acids Research, 2012, 1–7.
20. Honig, B., "Protein Folding: From the Levinthal Paradox to Structure Prediction J.Mol.Biol (1999) 293: 283-293.