2017 Biochemistry 110
California Institute of Technology
Lecture 3: The Structure of Myoglobin, and Bioinformatics

1825:
J. F. Englehard calculated the molecular weight of hemoglobin from the known atomic mass of iron: $\text{MW}_{\text{Hb}} = n \times 16000$.

1958:
John Kendrew (right) solved the X-ray structure of sperm whale Myoglobin.

1959:
Max Perutz solved the X-ray structure of horse Hemoglobin (Lecture 4).
(both achievements used the “heavy atom method” of introducing mercury ions selectively to the protein crystal in order to solve the phases of the diffraction data.)

Relevant Reading from Chapter 6: pp. 169-184.
Myoglobin has a Compact Structure with α-Helices

The sperm whale uses myoglobin to store oxygen for deep dives.

1957 Kendrew Structure: (6Å resolution)

1. extremely compact
2. ~ 75% α-helical (eight)
3. Some prolines appear to be placed as helix terminators.
4. the inside and outside of the protein are well-defined.

Val-Leu-Ser-Glu-Gly-Glu-Trp-Gln-Leu-Val- 10
Leu-His-Val-Trp-Ala-Lys-Val-Glu-Ala-Asp- 20
Val-Ala-Gly-His-Gly-Gln-Asp-Ile-Leu-Ile- 30
Arg-Leu-Phe-Lys-Ser-His-Pro-Glu-Thr-Leu- 40
Glu-Lys-Phe-Asp-Arg-Phe-Lys-His-Leu-Lys- 50
Thr-Glu-Ala-Glu-Met-Lys-Ala-Ser-Glu-Asp- 60
Leu-Lys-Lys-His-Gly-Val-Thr-Val-Leu-Thr- 70
Ala-Leu-Gly-Ala-Ile-Leu-Lys-Lys-Gly- 80
His-His-Glu-Ala-Glu-Lys-Pro-Leu-Ala- 90
Gln-Ser-His-Ala-Thr-Lys-His-Lys-Ile-Pro- 100
Ile-Lys-Tyr-Leu-Glu-Phe-Ile-Ser-Glu-Ala- 110
Ile-Ile-His-Val-Leu-His-Ser-Arg-His-Pro- 120
Gly-Asn-Phe-Gly-Ala-Asp-Ala-Gln-Gly-Ala- 130
Met-Asn-Lys-Ala-Leu-Glu-Leu-Phe-Arg-Lys- 140
Asp-Ile-Ala-Ala-Lys-Tyr-Lys-Glu-Leu-Gly- 150
Tyr-Gln-Gly 153
High-Resolution Structure of Myoglobin: (Kendrew’s structure from 1958, 1mbn)

Critical amino acid residues in myoglobin:
Proximal histidine (F8) and Distal Histidine (E7)
Phenylalanine (CD1), Leucine (F4) – heme contact
Glycine B6 – allows close approach of B and E hel.
Proline C2 – terminates the B helix and starts the C hel.

Val-Leu-Ser-Glu-Gly-Glu-Val-Trp-Gln-Leu-Val-NA1 NA2 A1 A2 A3 A4 A5 A6 A7 A8
Leu-His-Val-Trp-Ala-Lys-Val-Glu-Ala-Asp A9 A10 A11 A12 A13 A14 A15 A16 AB1 B1
Val-Ala-Gly-His-Gly-Gln-Asp-Ile-Leu-Ile-B2 B3 B4 B5 B6 B7 B8 B9 B10 B11
Arg-Leu-Phe-Lys-Ser-His-Pro-Glu-Thr-Leu-B12 B13 B14 B15 B16 C1 C2 C3 C4 C5
Glu-Lys-Phe-Asp-Arg-Phe-Lys-His-Leu-Gln-C6 C7 CD1 CD2 CD3 CD4 CD5 CD6 CD7 CD8
Thr-Glu-Ala-Glu-Met-Lys-Ala-Ser-Glu-Asp D1 D2 D3 D4 D5 D6 D7 E1 E2 E3
Leu-Lys-Lys-His-Gly-Val-Thr-Val-Leu-Thr-E4 E5 E6 E7 E8 E9 E10 E11 E12 E13
His-His-Glu-Ala-Glu-Leu-Pro-Leu-Ala-90 EF4 EF5 EF6 EF7 EF8 F1 F2 F3 F4 F5
Gln-Ser-His-Ala-Thr-Lys-His-Lys-Ile-Pro-100 F6 F7 F8 F9 FG1 FG2 FG3 FG4 FG5 G1
Ile-Lys-Tyr-Leu-Glu-Phe-Ile-Ser-Glu-Ala-110 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11
Ile-Ile-His-Val-Leu-His-Ser-Arg-His-Pro-120 G12 G13 G14 G15 G16 G17 G18 G19 GH1 GH2
Gly-Asn-Phe-Gly-Ala-Asp-Ala-Gln-Gly-Ala-130 GH3 GH4 GH5 GH6 H1 H2 H3 H4 H5 H6
Met-Asn-Lys-Ala-Leu-Glu-Leu-Phe-Arg-Lys-140 H7 H8 H9 H10 H11 H12 H13 H14 H15 H16
Tyr-Gln-Gly 153
HC3 HC4 HC5
Comparison of Sequences for Similarity/Homology

- Sequence Analysis suggests that Chopin’s 4th Ballade (1852) derives from Beethoven’s Appassionata (1838).

Using Sequence Analysis, we will compare Proteins for Similarity and Homology.

How Much Similarity Can We Expect from:

a) The SAME protein from two different species?
b) Two Homologous Subunits from a Multimeric Protein (e.g., a Heterodimer)?
c) Two Different Proteins with Similar Function/Mechanism?
d) Two Completely Different Proteins?
Myoglobin and Hemoglobin as a Model for Bioinformatics
Comparing Proteins with Similar Functions:

Another Example:

Hemoglobin (α chain)  Myoglobin  Leghemoglobin

Bovine ribonuclease  Human ribonuclease  Angiogenin
Comparing Proteins by Sequence Alignment

Hemoglobin

Myoglobin

VLSPADKTNKAAWGVKGAHAVGYGAEALERMFLSFSSTK
GLSEGEGWQLVLNVWGMKVFLDPHIGQPDVEILSLLKGGGHGHELE
YPPESDLHSHGSAQVHGKXGKVVAJLNAVANHVDMPNALSA
KFDKFKNLKSLDKEMKAG8DLKKGATLGLILKLKKHGGH
LSDLHAHKLRLVPNPKLLSHCOLVTLAAHLPAEPFPAVHA
EABIKPLAQSHTAKHPIPVXKYLEPISCECTIQVLQSKKHPDF
SLGKFLASVTVLTSSYR
GADAQGAMAMKALELRKDMASNYKXELGPQG

22 matches

VLSPADKTNKAAWGVKGAHAVGYGAEALERMFLSFSSTK
GLSEGEGWQLVLNVWGMKVFLDPHIGQPDVEILSLLKGGGHGHELE
YPPESDLHSHGSAQVHGKXGKVVAJLNAVANHVDMPNALSA
KFDKFKNLKSLDKEMKAG8DLKKGATLGLILKLKKHGGH
LSDLHAHKLRLVPNPKLLSHCOLVTLAAHLPAEPFPAVHA
EABIKPLAQSHTAKHPIPVXKYLEPISCECTIQVLQSKKHPDF
SLGKFLASVTVLTSSYR
GADAQGAMAMKALELRKDMASNYKXELGPQG

23 matches

Number of matches

Alignment
Introduction of Gaps for Sequence Homology Search

Using one gap, now we have 38 matches!

But are we just fooling ourselves?
We can randomize one of the sequences and produce 300 new sequences. These can be “scored” using a computer algorithm →
But the actual sequence has the highest score!
Scoring System for Finding Sequence Homology

Blosum-62 Scoring:

<table>
<thead>
<tr>
<th>Starting amino acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Human Myoglobin vs. Hemoglobin

Leghemoglobin vs. Myoglobin

(A) Alignment score (identities only)

(B) Alignment score (Blosum 62)
How Much Similarity Can We Expect from:

a) The SAME protein from two different species? Human and Chimp cytC differ by only 1 aa.

b) Two Homologous Subunits from a Multimeric Protein (e.g., a Heterodimer)?

→ Human Hb-α and Hb-β have 55 identities out of 141 aa using two gaps.

c) Two Different Proteins with Similar Function/Mechanism? – see ruvC and HIV integrase

d) Two Completely Different Proteins? note the randomizing strategy on previous slides

An Example Problem:

For the two protein sequences shown below, search for sequence alignment by offsetting each sequence by up to 10 residues relative to the other (see directions below). How many identities do your four best alignments afford? Allowing for the insertion of two gaps (each of which may be 1-10 residues long), show the best alignment that can be achieved. What is the most likely relationship of these two sequences?

a. They are sequences of two homologous subunits of a multimeric sheep protein.
b. They are sequences of the same protein, one from human and one from sheep.
c. The two sequences are of two completely unrelated proteins.

Directions: Using the Excel Spreadsheets pick one sequence and cut the entire sequence including the "-" marks on either side and paste it back in offset by +1 to +10 units. Record the # identities for each offset and inspect to see if these are clustered. Then do the same for the other sequence. Then allow yourself two gaps (that you will fill in with "-" marks) and try to produce a single alignment with the greatest number of identities.

SEQUENCE A:
MRECISIHVGQAGVIGNACWELYCLEHGIQPDPQMPDSKTIGGGDSFNTFFSETGAGKHWAPRVFVDLEPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIDLVDRIRKADQCTGLQFLGVFHSFGGTSGFTSLLMERLSDYKSKLEFSIYPAPQVSTAVVEPYNKTSILRTHTTEHSDCAFMVDNAAIDRNLDIERPTRYTLNLRLIGQIVVSSSTASLRFDGALNVDLTEFQTNLVPRHPHPLATYAPVISAEGAYEQSLVAETNACFEPANQVNCPRGKYMCHELLLHRGDVLPVDVNAATIKTRTIQFVDCPTGKVQNYQPPTVVPGDGLAKVQRACMLSNTEAIAEAWORLHFDLMYAKRAFVWHYVYVEKMEEGFSEAREMDMAALEKDYEEVGDSVEGEHGEEEDEEY

SEQUENCE B:
MREVHIOAQQCNQIGAKFWEVISDEHGIIDPTGYSYMGDSQDLQLEINVVYNEATGNGKVPRAILVDEPQMDTCSRSGPFGQIFRPDNDVFQGQSSAGNNWAKGHYTEGAEOVSVDLVRKLESCECDQCQFQLTHSLGGTSGGMTLISKIREYPRDIMNTFSVMSKVPSTVVEVYNATLSVHQLVENTDETYSIDEALYDICFRTLKLTTPYGDNLHVSATMSGVTTCLRFPGQNLADRKLAVNMVPFLTHFMPGFAPLTSRGSSQYRALTVPELTQOMFSKNNMAACDPHRGRYETLVAWAVPGRMSMKEVDEQMLNVQNKNSYFVWEIIPNNVTAKCDIPRGLKMSATFIGNSIAIQELFKRISEQFTAMFRRKAFLHYWTGEMDEMERTEAESNMDLQVSEYQQYQDATADEQGEFEEEGEEDEA
Example Solution

Search for sequence alignment by offsetting each sequence by up to 10 residues. How many identities do your four best alignments afford? 41 (0), 142 (+2), 42 (+7) and 53 (+10). Allowing for the insertion of two gaps, show the best alignment that can be achieved. (How many identities does this alignment effort afford?). 180 identities out of 451 aa (40% identities):

SEQUENCE A (sheep α-tubulin, top) SEQUENCE B (sheep β-tubulin, bottom):
If they were the same tubulin subunit from human/sheep, they would be >>40% homologous (actually 97%).

Nature Displays a Diverse Array of Sequences for each given Function

- Cytochrome c performs an essential role in oxidative phosphorylation.

The Dickerson Mosaic – built in 1974 from ¾” colored tiles by the “Caltech Cytochrome Seers”

<table>
<thead>
<tr>
<th>Human, chimp:</th>
<th>GDVEGKKKIFIMKCSQCHTVEKGGHKTGPNLHGLFGRSOGAPYSYTAANKGIIWG EDTLMEYLENPKKYIPGKIMFVGIKKEERADL...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhesus monkey:</td>
<td>G----G---F---C--CH----------GP-L-G---R--Y---AN------W--YL-NFKKYIPGKIMFVGIKKEERADL...</td>
</tr>
<tr>
<td>Horse:</td>
<td>G---G---C---CH--------------GP-L-G---R--G--G--Y--AN------W--YL-NFKKYIPGKIMFVGIKKEERADL...</td>
</tr>
<tr>
<td>Donkey:</td>
<td>G---G---C---CH--------------GP-L-G---R--G--G--Y--AN------W--YL-NFKKYIPGKIMFVGIKKEERADL...</td>
</tr>
<tr>
<td>Cow, pig, sheep:</td>
<td>G---G---C---CH--------------GP-L-G---R--G--G--Y--AN------W--YL-NFKKYIPGKIMFVGIKKEERADL...</td>
</tr>
</tbody>
</table>

Moths, flies...

yeasts, molds...

wheat, beans, pumpkin, sunflower,...

Heme binding
### Cytochrome c Enzymes Compared by Species

<table>
<thead>
<tr>
<th>Species</th>
<th>Average Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man</td>
<td>0</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td>1</td>
</tr>
<tr>
<td>Horse</td>
<td>12</td>
</tr>
<tr>
<td>Donkey</td>
<td>11</td>
</tr>
<tr>
<td>Pig, Cow, Sheep</td>
<td>10</td>
</tr>
<tr>
<td>Dog</td>
<td>10</td>
</tr>
<tr>
<td>Gray whale</td>
<td>9</td>
</tr>
<tr>
<td>Rabbit</td>
<td>9</td>
</tr>
<tr>
<td>Kangaroo</td>
<td>9</td>
</tr>
<tr>
<td>Chicken, Turkey</td>
<td>8</td>
</tr>
<tr>
<td>Penguin</td>
<td>8</td>
</tr>
<tr>
<td>Pekin duck</td>
<td>8</td>
</tr>
<tr>
<td>Rattlesnake</td>
<td>8</td>
</tr>
<tr>
<td>Snapping turtle</td>
<td>8</td>
</tr>
<tr>
<td>Bullfrog</td>
<td>8</td>
</tr>
<tr>
<td>Tuna fish</td>
<td>8</td>
</tr>
<tr>
<td>Screw-worm fly</td>
<td>8</td>
</tr>
<tr>
<td>Silkworm moth</td>
<td>8</td>
</tr>
<tr>
<td>Wheat</td>
<td>8</td>
</tr>
<tr>
<td>Neurospora crassa</td>
<td>8</td>
</tr>
<tr>
<td>Baker’s yeast</td>
<td>8</td>
</tr>
<tr>
<td>Candida krusei</td>
<td>8</td>
</tr>
</tbody>
</table>

**Comparison of Cytochrome c**

**CytC Universally Conserved aa’s**
Genetic Recombination and the Positions of Genes

Human chromosomes

Mouse chromosomes
Recombination and the Role of the Holliday Junction

1964: Discovered by Robin Holliday (1932-2014)

Double-strand break

\[ \text{Resection} \]

3' \hspace{1cm} 5'

\[ \text{Stand invasion, D-loop formation, DNA synthesis} \]

DSBR

SDSA

Second end capture, DNA synthesis, Ligation

\[ \text{Branch migration, Resolution} \]

Non-crossover (uncommon)

or

Crossover (common)

Strand displacement annealing

DNA synthesis, Ligation

Non-crossover
The Holliday Junction Resolving Enzyme Complex has Three Multimeric Subunits

ruvA tetramer with DNA (1BDX), top view:  ruvA tetramer with DNA (1BDX), side view:
HIV Integrase (a viral homodimeric enzyme that inserts viral DNA into the Host genomic DNA) Shares Homology with HJ Resolving Enzyme ruvC Subunit (homodimeric structure shown top left - essential in recombination)