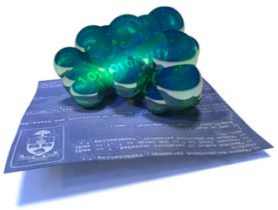
A
Bioinformatics
Course

Alignment

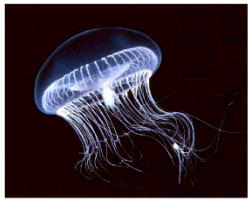


Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

similarity

Homologous Proteins: Conserved structure and function



Green Fluorescent Protein
(*Aequorea victoria, 1EMA*)

Red Fluorescent Protein
(*Discosoma striata, 1GGX*)

```
GFP   MGKGEELFTGVVPILVELDGDVNGHKFSV
RFP   MRSSKNVIKEFMRFKVRMEGTVNGHEFEI

GFP   SGEGEGDATYGKLTLKFICTT.GKLPVPW
RFP   EGEGEGRPYEGHNTVKLKVTKGGPLPFAW

GFP   PTLVTTFSYGVQCFSRYPDHMKRHDFFKS
RFP   DILSPQFQYGSKVYVKHPADI..PDYKL

GFP   AMPEGYVQERTIFFKDDGNYKTRAEVKFE
RFP   SFPEGFKWERVMNFEDGGVVTVTQDSSLQ

GFP   GDTLVNRIELKGIDFKEDGNILGHK.LEY
RFP   DGCFIYKVKFIGVNFPSDGPVMQKKTMGW

GFP   NYNSHNVYIMADKQKNGIKVNFKIRHNIE
RFP   EASTERLYPRDGVLKGEIHKALKLK....

GFP   DGSVQLADHYQQNTPIGDGPVLLPDNHYL
RFP   DGGHYLVEFKSIY..MAKKPVQLPGYYYV

GFP   STQSALSKDPNEKRDHMVLLEFVTAAGIT
RFP   DSKLDITSH....NEDYTIVEQYERTEGR

GFP   HGMDELY
RFP   ...HHLF
```

Measuring similarity requires an **Alignment**. Calculating an alignment means accounting for **amino acid similarity, insertions and deletions**.

In order to infer homology we must measure similarity.

One such measure is the fraction of identical residues in two **aligned** sequences.

Obviously, the fraction of identical residues depends on the alignment and that raises the questions how we can obtain a correct alignment. But even before we can start aligning, we need to define a metric for amino acid similarity, because the right alignment should give us good **similarity**, not just a large percentage of **identical** residues. Also, we would like to have a measure that tells us how likely it is that the similarity in an alignment is due to evolutionary descent. And there is an additional issue: how do we treat sequence insertions resp. deletions in the alignment quantitatively?

alignment

An alignment is a notation for a mapping between a pair of amino acids. Amino acids that are in the same position of the alignment are mapped to each other to express some relationship. What exactly is the relationship that we express in biological alignments?

Consider: aligning a segment of GFP and RFP with unequal length.

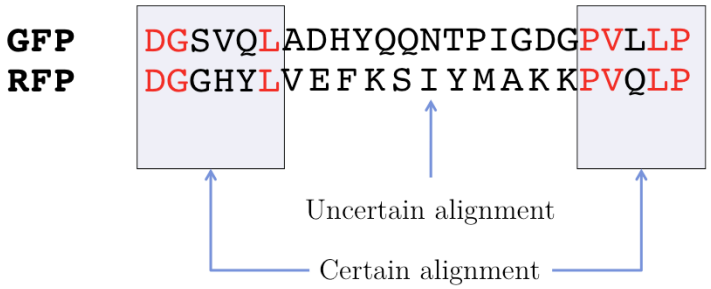
```
GFP   DGSVQLADHYQQNTPIGDGPVLLP
RFP   DGGHYL VEFKSIYMAKK PVQLP
```

We produce an alignment so as to maximize the similarity between amino acids in corresponding positions.

alignment

An alignment is a notation for a mapping between a pair of amino acids. Amino acids that are in the same position of the alignment are mapped to each other to express some relationship. What exactly is the relationship that we express in biological alignments?

Consider: aligning a segment of GFP and RFP with unequal length.



Alignments do not simply consist in writing one sequence above the other. An alignment is a map of correspondences and we need to find the correct alignment, that makes the correspondence meaningful. We want to be able to interpret the matched amino acids as a statement about the underlying biology: the pair of amino acids in an aligned position should be descended from a unique common ancestor.

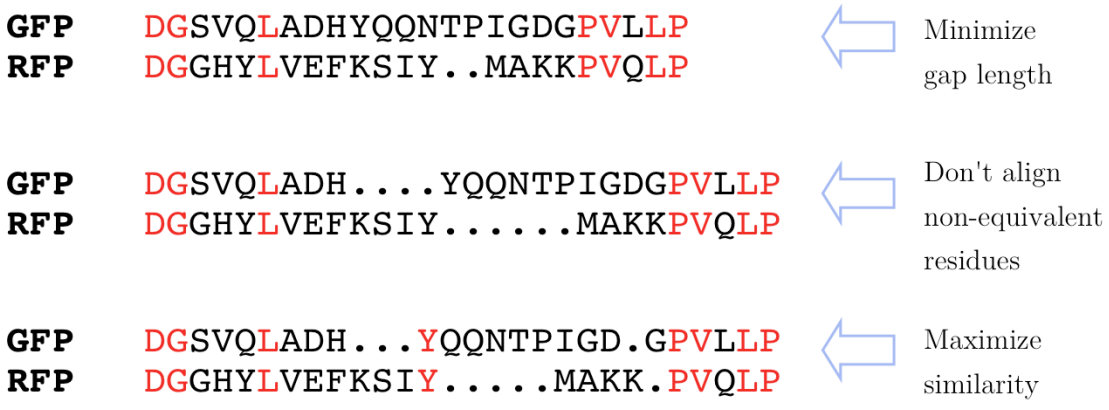
In some regions of the alignment (grey boxes), aligning for maximal pairwise identity is straightforward. There is only one, obvious way how to do that. But in other regions, there may be no uniquely best alignment, or the sequences may have different lengths.

Proteins evolve to have different lengths through changes at their N- and C-terminus, and internal insertions and deletions (indels). These length changes need to be reconstructed in order to produce an alignment. We need to figure out **where** to acommodate the indels.

alignment

In order for an alignment to make sense, we should strive not to pair-up amino acids that can not be compared on equal terms because they evolve in a very different structural context. But insertions/deletions always change the context over an unpredictable stretch of residues.

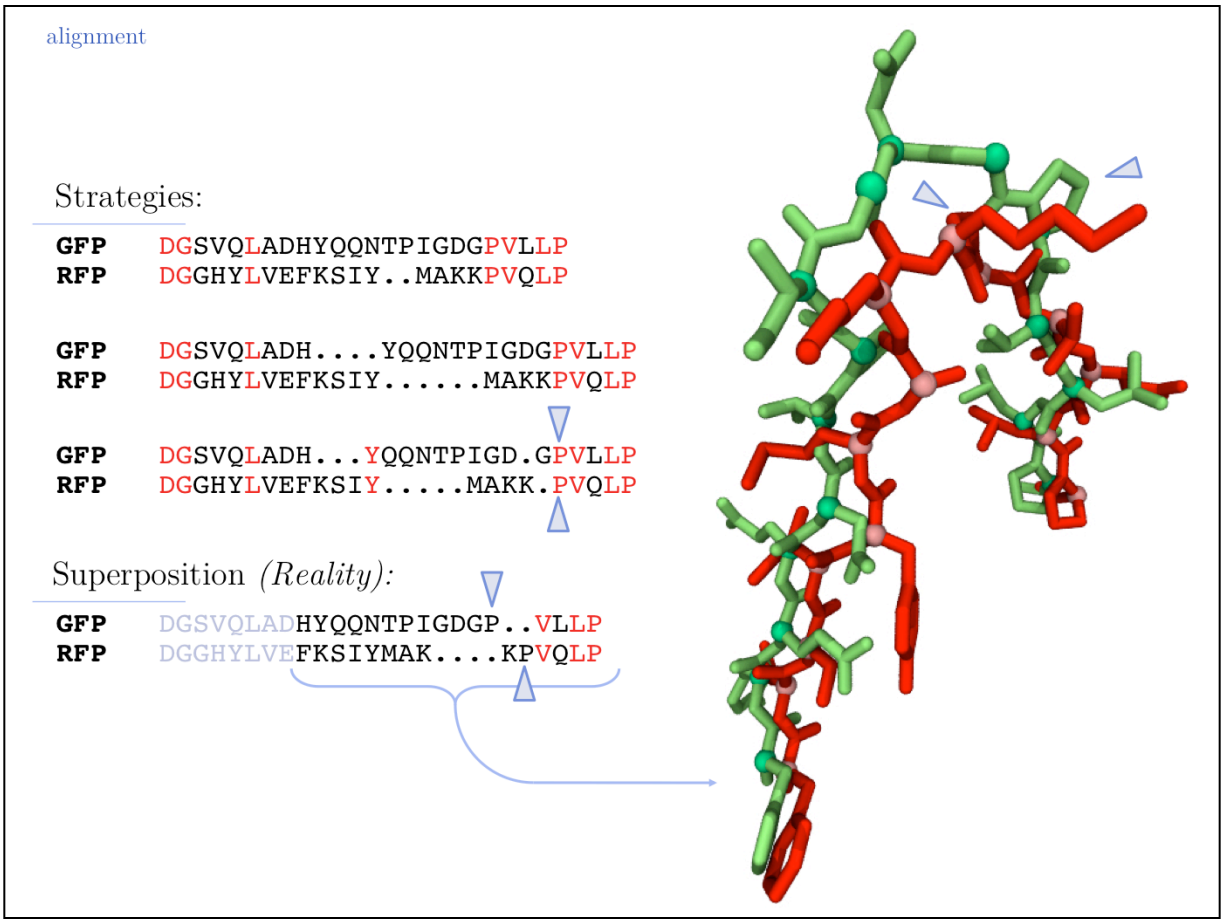
Strategies to resolve indels:



Merely *minimizing gap length* does not tell us **where** to place the indel.

Not aligning non-equivalent residues is the conceptually cleanest solution, but it produces alignments that are not compact and may miss important relationships.

Maximizing similarity may align residues that are identical, but not actually related.



Moreover, it's not clear what the **correct** alignment should be in the first place. We can consider a structure superposition to be something like the "ground truth" for sequence similarity, it captures the context in which each amino acid performs its function and experiences its selective constraints.

But the superposition does not necessarily capture the **historical process** of how a particular sequence change was generated and acommodated in evolution. Moreover, it does not necessarily correspond to any of the alignment heuristics we mentioned above. In the image above an alignemnt has been derived from the structural superposition of green- and red- fluorescent protein and residues that are structurally in a different context have been paired with hyphens. Note that the two prolines at the right hand aligned block that all of our heuristics had aligned, actually are **not** superimposed!

Part of the problem is that the structural accommodation of an indel is not necessarily the site at which the indel arose during evolution of the sequence.

Actual alignment algorithms don't really take this into account.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA