Protein Structure, Function and Evolution: Background and sequence analysis

Choose your own adventure

The Impact of Structural Genomics: Expectations and Outcomes

John-Marc Chandonia and Steven E. Brenner
SCIENCE 311 (20 JANUARY 2006), 347-351.

Background SG

- PSI projects move from pilot to full production phase
- Future budget: \( \approx 600M \) (US) annually
- Opportune time to examine results
  - “We quantitatively analyzed the novelty, cost, and impact of structures solved by SG centers, and we contrast these results with traditional structural biology”
- Helpful for informing future strategies
  - (analysis biased towards novel structures, the declared outset of SG)

New structures solved

Impact of SG on coverage of protein families

- SG contribute \( \approx 20\% \) of total structures but account for \( 50\% \) of new structures
  - New Pfam:
    - \( 5\% \) of non-SG structures
    - \( 20\% \) of SG structures
  - PSI = 2/3 of all SG
**Cost balance**

- In 2003 average structure under R0 grant was estimated $250k-$300k
- PSI average cost: $138k (46-59% of non-SG)
- Normalised per residue (higher degree of difficulty with larger proteins):
  - SG cost = 66-85% of non-SG

**Summary**

- Publication is a bottleneck in SG
- Not easily adapted to high through-put
- All these analyses should be taken with extreme caution
- Time lag due to startup
How to characterise a protein?

- Structure
  - Related structures may have related function
- Evolution
  - Functional residues are conserved
  - Function may be different in different organisms (environments)

Sequence collection and alignment

Sequence collection

- Many methods for structure and function prediction, and all evolutionary methods, rely on having not just one sequence, but many related sequences in a multiple sequence alignment
- Collect related sequences using tools like BLAST
- Make an alignment using tools like clustalX

Blast searches

- I recommend looking for related sequences mainly at the protein level: blastP
- The DNA level is mainly useful for very similar sequences
- Also try PSI-Blast (based on blastP) to find more distantly related sequences.
- You can also search profile databases eg pfam

PSI-blast

- PSI-blast works using multiple interactive steps.
- The first round is normal blastP
- Related matches found in the first round are incorporated into a profile.
- The next round uses the profile to search, finding more distant relatives.
- When running PSI-blast on the command line, you specify a significance cut-off and the number of rounds to do. You will probably want to experiment with the number of rounds (start small) and the significance cut-off to avoid unrelated sequences becoming part of the profile.
Psi-Blast

Editing to region of similarity

- Often, blast matches are longer than the region of actual match. You need to cut out the extra sequence, and gather together only the similar regions that can be aligned.
- E.g. genomic sequence, match to a single gene, or multi-domain proteins matching to a single domain.

Fasta format

Fasta format is commonly used for sequences. It is easily human and computer-readable

- Hint: don't save sequences as Word docs. Use plain text.

Keeping track of it all...

- I suggest making a database or two (e.g. table on your wiki) including:
  - A short name for the sequence
  - Database details like accession number
  - Species and other taxonomy information
  - Sequence length, and region that you are using, if not the full-length sequence
  - any other notes

Constructing a multiple sequence alignment

- It's not enough to have a collection of similar sequences - we need to align them to highlight the pattern of similarity.
- Use a program like clustalX or muscle
- Make an input file for clustalX: Use fasta format.
- Edit the sequences if necessary to correspond to your sequence. E.g. DNA blast matches may be from genomic sequence, so look for the CD region that encodes the actual protein.

Running ClustalX

- Once you have loaded the sequences into ClustalX, from the menu “Align” choose “do complete alignment now”.
- ClustalX will report progress in the bottom left hand corner
- If the alignment is running really slowly or seems to get “stuck” at a particular point, the most likely reason is that some of the sequences you have included don’t match others
**What ClustalX is doing**

- First, all input sequence are compared pairwise.

- It then finds the sequences that are most similar to each other to align first, adding the less similar sequences later.

- Finally it builds the multiple sequence alignment based on the order it found.

**ClustalX output**

- Once clustalX has finished, you should look carefully at the result.

- The sequences should look aligned to each other!

- You can select different output formats. It may be worth saving your alignment in different formats as different programs prefer different formats. In particular, .phy is the phylip alignment format.

- You can look at the .dnd file to see the alignment order that was chosen.

**Improving the alignment**

- You may need to edit your original sequences and import them to clustalX to align again a few times until the alignment looks good.

- Check whether some sequences are much longer than others (ragged ends), and whether there are long insertions in some sequences, and edit the sequences accordingly.

- Remember to make a record of the editing you do.

- You may need to change alignment parameters for your sequences: for example if you have a lot of small gaps in one sequence.

**Changing gap costs**

- Gap cost: 8 opening + 7 extension

  GTTCGTAGAGTTAAGAGAGCGTTTCGAATCAGTAAGAG

  GTT---T---G---T---A---GAG----TTT---AAG---AGTAAGAG

- Gap cost: 1 opening + 14 extension

  GTTCGTAGAGTTAAGAGAGCGTTTCGAATCAGTAAGAG

  GTTTGTAAGATTGAAG------------------------AGTAAGAG

- So, by increasing the cost of opening a gap and reducing the extension cost, you can change the alignment.
Judging the alignment

- Every alignment is different, but there are some things to look for.
- In a DNA alignment, if the DNA is protein coding, gaps should be multiples of three, and the most variable positions should be the third position in the codon.
- In an amino acid alignment, most positions should have conservative amino acid changes.
- If there’s a known structure, conserved parts of the structure should have fewer gaps.

Amino acid similarities

- Try to become familiar with amino acid properties and similarities, e.g., photocopy reference pages from a textbook.
- Examples: L, I, V (leucine, isoleucine, valine) hydrophobic, about the same size.
- D, E (aspartate, glutamate) acidic
- R, K, H (arginine, lysine, histidine) basic
- C (cysteine) forms cysteine bridges
- ClustalX colour coding is useful

If things look bad...

- If one sequence isn’t aligning with the rest, check that it really matches in your original blast.
- If it’s DNA, check that it’s not reverse complemented.
- You can use “blast two sequences” to compare the trouble-maker to a well-behaving sequence.
- Check the literature: has someone already dealt with this problem?
- Get a tutor to help.

If several regions match

- Your blast results may not be full length matches but lots of partials:
- In this situation, you need to make two alignments.
- Focus on quality matches (red and pink).

Alignment program limitation

- ClustalX will not be able to align sequences that don’t align to each other.
- Most MSA programs will find this difficult, at least.
Summary

- Start with your sequence of interest
- Find related sequences using blast - think about which blast and which database is relevant
- Construct a multiple sequence alignment e.g. using ClustalX
- Keep records of the sequences and your editing
- Be prepared to trouble-shoot and have to repeat alignments
- The multiple sequence alignment will be useful for both evolutionary and structural analysis.