Programming for Bioinformatics - part 2

Saul Pierotti

January 10, 2020

Intro

- Pseudocode is a script in natural language
- Homology is boolean: 2 sequences are either homolgs or not
- I can do homology building only if I have homologs (!)
- Python is read line by line, so I cannot define a function after calling it (!)
- Code of conduct were born in programming communities

Structural alignment

- The superimposition problem can be formalized as, given 2 sets of points A and B, finding the optimal subsets of A and B for which the norm of the 2 subsets is equal and the optimal rigid body trasformation such that minimizes a distance metric
- The RMSD is the average distance among alpha carbons
  - It is calculated as the sum of squares of the coordinates, under square root, divided by the number of points
    - The square of the coordinates under square root is the pitagorean theorem in 3 dimensions (!)
- Over 60% identity I have more than 90% of residues with RMSD less than 1 Å

Sequence alignment

- The score of a position can be considered independent from any other position
- It is not always possible to distinguish spurious alignments from true ones
- If we assume independence, the probability of having a string of nucleotides is the product of the probability for each nucleotide in the string
- If we want to be able to sum scores, we need to take the log of the probabilities (!)
- In RNA indipendence is not verified: a position influences the likelyhood of the one next to it
  - High significant scores may be spurious (!)
- The raw score of an alignment is the sum of the score of each match plus the sum of the score of the gaps
- For gaps I usually have an open gap penalty and an extension penalty
  - If I work on the core, I normally use the same open and extend penalties
- Similarity takes into account the physicochemical properties of residues
- Mutations (expecially indels) in the core are rare because the protein could completely lose structure
- Loops tend to mutete frequently
- A scoring scheme defines a distance between sequences
- Alignment algorithms can be exhaustive (slow) or euristic (fast)
- 2 every 3 SNPs are transitions
- When computing a log-odds, it can happen that the observed frequency of a substitution is 0
  - In this case, we want to add 1 to any count, so to be able to take the logarithm
  - We add 1 to ALL the scores so that we do not introduce a bias

  - This is called pseudocount
- Homology is boolean: 2 sequences are either homolgs or not
- Scoring matrices can be based on observed substitution, on physicochemical properties, or other data
- The McLachlan matrix is based on residue similarity, Grantham on chemical distance

PAM matrices

- PAM matrices are based on an evolutionary model, while BLOSUM are based on real alignments
- PAM were created by M. Dayhoff in 1972 and their name is an acronim for "point accepted mutation"
- A point accepted mutation is the replacement of a residue by another, which is accepted by natural selection
- The matrix PAMx is a matrix referring to sequences undergoing x PAM every 100 residues
  - Note that this includes multiple mutations at the same site (!)
- The protein set chosen for building PAM matrices had a minumum identity of 85%
- From the set of proteins, they inferred a phylogenetic tree
  - The tree is used to follow all the chain of mutations, also the ones that happened in the same positions
- Higher PAM matrices are computed as powers of PAM1
  - This means that we are using a model, it is not based on observation (!)
- PAM1 is a really stringent matrix: all the values outside the diagonal are really low
- The scores of PAM1 are computed as log_odds of the substitutions

BLOSUM

- Blosum means block substitution matrix
- They were built by Henikoff and Henikoff in 1992
- They started from 500 multiple local alignments or related protein sequences
- From these they retrieved more than 2000 ungapped multiple sequence alignments (conserved blocks), that formed the BLOCKS database
- In each multiple alignment, they clustered all the sequences that showed a % identity and average their log-odds
- They were made by alignments of protein sets with different degrees of conservation
  - We do not assume any model, all the substitutions are based on observed frequencies
- High BLOSUM matrices are used for closely related sequences
- We can build our own BLOSUMs based on a specific subset of proteins
- BLOSUM62 is the best performing matrix for comparing weak homologies