

Laboratory of Bioinformatics 1 part B - Capriotti

Saul Pierotti

April 2, 2020

Introduction

- There will be a project for the final, to be submitted for May 18
- Protein structure is more conserved than sequence
- When sequence identity is sufficiently high, we can transfer structural information
- A structural alignment is a rigid body transformation of 2 subsets from 2 sets of points that maximizes a given distance metric
 - The subsets need to have the same number of elements and define the correspondence set
 - Finding the correspondence set is an NP-hard problem
 - Finding the optimal rigid transformation of the correspondence set is $\Theta(n)$
- The distance of 2 sequences can be evaluated with a substitution matrix and a gap penalty
- Global alignments are computed with the NW algorithm, local alignments with the SW
- Local alignments are useful for multidomain proteins, when only some domains are conserved
- The significance of an alignment score can be evaluated by comparing with the score distribution for random alignments
- Sanders and Schneider developed a twilight curve before Rost
 - Sander's curve becomes straight after 80 residues
 - Rost's curve does never become straight
- Over 100 residues, under 25% of sequence identity only 10% of the sequences are homologous, while above 20% 90% of them are
- Over 30% identity sequences longer than 100 residues have similar structures, but this does NOT mean that under 30% the structure is necessarily different!
- Proteins with low sequence identity but high structural similarity are referred to as remote homologs
- Structures can be predicted by comparative modeling, threading, ab initio
- If I want to use sequence identity for transferring annotation features, I need to identify the problem-specific twilight region
 - For subcellular localization, the twilight zone is 50% (!)
- The sequence identity needed for transferring subcellular localization is higher than that required for structure
- Function of proteins with really high sequence identity can be completely different
- In remote homologs the sequence alignment is often wrong
- Important residues in a sequence can be identified by comparing conservation levels

Structural alignment

- Structural alignment is different from superimposition
- Superimposition assumes that I already have the correspondence set, and it is relatively easy
- Structural alignment requires the identification of the correspondence set, which is hard
- The definition of domain is often heuristic and questionable
- Proteins with similar spatial distribution but different topology are difficult to align
- Alignment methods can be classified in different ways

- Pairwise or multiple
- Depending on the descriptor used
 - * Backbone
 - * All atoms
 - * Sequence-based
 - * Contact map
 - * Surface
- Rigid body or flexible
- The comparison of torsion angles is $\Theta(n)$
 - They are invariant for rotation and translation
 - It is good for local regions but problematic for whole structures
- A distance matrix is also invariant for rotation and translation
 - Comparing matrices is hard, $\Theta(n^2)$
 - It is not sensitive to chirality
- At the moment, all methods are able to identify obvious similarities
- Remote similarities are detected by a subset of methods, and different methods recognize different similarities
- Speed is an issue in many algorithms
- We want our method to be biologically meaningful, not only geometrically
- The expected score or random pairwise alignments is an extreme value distribution
 - I would have a gaussian if there was no evolution
 - In real databases I have an excess of good-scoring pairs
- When I want to determine the distribution of scores, it is better to have an analytical distribution than an empirical one
 - I don't have tools for working with empirical distributions (!)

CE algorithm

- Compares AFPs composed of 8 residues, stitches them together and finds an optimal path through them with dynamic programming
- It gives a statistical score
- The alignment is the longest continuous path of AFPs in a similarity matrix S
- The similarity matrix S is composed represent all AFPs conforming to a similarity criterion
- The dimensions of S are $(n_a - m)(n_b - m)$, where n_a and n_b are the length of the sequences and m the size of the AFPs
- The matrix is large to compute, therefore we need constraints
- Two consecutive AFPs can be aligned with a gap in protein A, a gap in protein B or without gaps
- The AFP length is set to 6 and the maximum possible gap to 30
- Similarity measures are RMSD, full set of distances, and others
- The best 20 alignments with Z score above 3.5 are compared based on RMSD and the best one is kept
 - I get an error in 1000 comparisons
- Each gap is assessed for relocation up to $m/2$ times
- Iterative optimization with dynamic programming
- It cannot find non-topological alignments
- The unit of comparison was originally the protein chain, but domains are optimal
 - Domains are difficult to define (!)
- The statistical distribution of alignment scores can be used to evaluate the Z score of an alignment

PDBe Fold

- It uses secondary structure elements (SSEs)
- Secondary structure is typically conserved
- SSE are represented as vectors that connected in a graph by edges
 - 2 vertices and an edge describe position and orientation of the SSEs

- SSEs are helices and strands
- Each edge is labelled by a property vector containing information on edge-vertices angles, torsion angles between vertices, length of the edge
- The set of vertices, edges and labels defines the graph that is then matched with an algorithm
- Vertex and edge lengths are compared both in absolute and relative terms
 - In relative terms, the same absolute difference is less significant for longer edges
- Torsion angles are used for distinguishing mirror symmetries
- The SSE matching gives correspondences among SSEs, and can be used to yield an initial sequence alignment
- Connectivity (topology) can be neglected, considered but allow for any number of missing SSEs (soft connectivity) or allow only for an equal number of unmatched SSEs (strict connectivity)

MAMMOTH algorithm

- Matching molecular models obtained from theory (MAMMOTH) is one of the fastest algorithms
- The protein is represented as a set of unit vectors among Ca
- It is based on dynamic programming
- An unit vector is the normalized vector among Ca atoms
 - For each position, k consecutive vectors are mapped into a unit sphere that represents the local structure of k residues
- Each set of unit vectors is compared to all the sets in the other structure, building a matrix
- Each comparison yields a unit root mean square distance (URMS)
 - This is compared against the expected random URMS
 - The alignment score is obtained by normalizing the URMS with its expected value
- The path through the matrix is found with dynamic programming by a global alignment without end-gap penalties

RNA structure

- Most RNAs are around 50 bp
- Secondary structure of RNAs is usually represented with parentheses
 - I cannot represent pseudo-knots in this way
- For RNA, the secondary structure is much more informative than for proteins
 - A certain secondary structure constraints a lot the tertiary structure
- There is less variability in RNA structures than in proteins
- The best atom for representing the backbone is C3', since it has the most constant inter-nucleotide distance
- The professor adapted MAMMOTH to work with RNA C3' atoms instead of Ca in proteins: SARA
 - The statistics of the score had to be re-evaluated
 - They still used the extreme value distribution, which is defined by μ and σ
 - They selected how the parameters change when RNA size changes
 - The set of unit vectors was 3 instead of 7
 - The method gives a $-\log(\text{p-value})$ score
 - By comparing RNAs of known function, I can determine a score threshold that gives correct functional annotation
- Another method was developed in Israel: ARTS
- Few people are working in RNA: not so many methods
- The twilight zone of RNA sequence alignment is around 60%
- Secondary structure identity (PSS) correlates well with tertiary structure identity (PSI) but not with sequence identity

Multiple sequence alignment

- In MSA it is easier than in pairwise alignments to identify conserved regions, that could be functionally important
 - We can observe blocks of conservation in MSAs
- I can transform a MSA in a profile of the sequences
- A profile is a matrix with a row for each possible residue and a column for each position
 - The value of each element reflects the frequency of a residue in a specific position
 - Each position is therefore a vector of 20 elements
 - I represent a profile as a matrix containing as many vectors as the number of positions
 - I can also have a row for the presence of a gap in the position
- A sequence logo is a plot showing the entropy of each residue in each position
 - It is obtained from a profile and it is a way to represent it
- Shannon entropy: information content of a message
 - For a single column $S(p) = \sum_{i=1}^{20} -p_i \ln p_i$
 - Total conservation: $S(p) = 0$
 - All residues are equally probable: $S(p) = \ln(20)$
 - There are more sophisticated models that take into account the expected frequency of residues
 - The entropy of an alignment is obtained by summing the Shannon entropy over the all alignment
- Scoring an MSA: sum of pairwise scores or entropy score
 - Not all the positions are equal in an MSA: some conservations are critical, others not
 - Scoring has necessarily to depend on the evolutionary history of the sequences
 - Almost all scoring functions assume positional independence
- I can score each pairwise alignment and sum it
 - $S = \sum_{i < j} S(A_i, A_j)$
- I can score an MSA depending on its entropy
 - The best alignment is the one with the lowest entropy (i.e. the most conserved one)
 - It is the sum over the alignment of the entropy in each position
 - $S = \sum_{j=1}^{N_{cols}} \sum_{i=1}^{20} -p_i \ln p_i$
- I can align a sequence to a profile
 - Each position is aligned to a vector for the position
 - The score for the position of the residue in the sequence with every possible residue is summed and weighted for the frequency encoded in the vector
 - * This is a matrix by vector multiplication (!)
 - These scores can be used with a dynamic programming algorithm

Algorithms for MSAs

- Dynamic programming approaches exist, but they are $O(N^M)$ and they are np-hard
- An MSA method can be evaluated from the functionally important residues that are correctly aligned

Progressive MSA

- ClustalW is an example of progressive MSA
- I align sequences in pairs, one after the other
- The result depends on the order of how I pair sequences (!)
 - I usually pair the most similar sequences first
- Similarity is measured by Kimura distance (see MUSCLE for more info)
- From each pairwise alignment, I build a profile
- I iterate until there are no sequences left, by aligning pairwise sequences and profiles
- In order to do this I need to be able to align profiles (!)
- I want to be conservative with gaps with the initial pairwise alignments, and introduce them later on profiles
 - When I get to profiles I have info about conservation (!)

- Errors in the first alignments are propagated
 - If I am not conservative I can become full of gaps
- I can improve the alignment by changing the sequence tree
 - By default Clustal uses NJ
 - Maybe I have a tree available (!)
- Adding gaps is tricky, since their penalty logically depends on the position and conservation
 - They are usually added in the first alignments
- In ClustalW the penalty is multiplied by a factor which is context specific
 - Gaps in hydrophobic regions are more penalised
 - These coefficients were derived from gaps frequencies in a large number of structural alignments
 - Gaps are discouraged if there is another gap nearby in the MSA
- Low-scoring alignments are postponed for later by adjusting the tree
 - ClustalW aligns them when it has more information deriving from the profiles
- A profile-to-profile alignments involve the pairwise comparison of same-dimensional vectors
 - I do a double sum all against all elements weighted with a substitution matrix
 - This is done via a simple vector to matrix multiplication, followed by a multiplication for the remaining vector (!)
- ClustalW corrects for biased representation of subfamilies
- The scoring matrices used change depending on the similarity of the sequences to be compared
- In general, ClustalW uses an heavily crafted heuristics
- The main problem of progressive alignment: subalignments are frozen in place
 - Once aligned, a group of sequences cannot be re-aligned by taking advantage of the new information deriving from other sequences

Iterative MSA

- Iterative MSA tries to overcome the problem of frozen subalignments
- MUSCLE: multiple sequence comparison by log expectations
- It is based on 3 steps: draft progressive, improved progressive, and refinement
- Draft progressive: create a first progressive MSA
 - Sequence similarity is defined by k-mer distance, not pairwise alignment score
 - * If a rare kmer is present in 2 sequences maybe they are related
 - It creates a distance matrix with all sequences against each other
 - It uses UPGMA instead of NJ for building the tree from the matrix
 - The score is based on log expectations, not pairwise score for profile to profile alignments
 - * It is the entropy score
- Improved progressive: from the draft create a new matrix and from that a new tree and a new alignment
 - The pairwise distances are calculated from the Kimura distance
 - * $K_{dist} = -\ln 1 - D - D^2/5$, where D is the pairwise identity
- Refinement: cut and re-align the tree
 - 1 edge is deleted at random from the tree
 - The 2 resulting profiles are re-aligned to each other to get the full MSA
 - If the score improves, keep the new MSA otherwise keep the previous one
 - This is iterated until convergence on a local minimum

Consistency-based MSA

- Consistency: if residue X is aligned with Y and Y is aligned with Z, then X is aligned to Z
 - This is necessarily true in an MSA
- In reverse, I can use consistency to align two sub-alignments: I take advantage of transitivity of alignments
- MSA are not necessarily consistent with the respective pairwise alignments
 - Progressive MSA methods frequently are not consistent with the pairwise alignments used for building the tree

- T-Coffe (tree-based consistency objective function for alignment evaluation) is an MSA method based on consistency
 - Build the primary library
 - * I do all the possible pairwise alignments and I measure the pairwise sequence identity
 - * Each pairwise alignment is equipped with a weight equal to the average identity of matched residues, ignoring gaps
 - Build the extended library
 - * In order to align sequences A and B, I try all the possible alignment, direct and based on an intermediate sequence C
 - * The weight of each alignment is the minimum of the pairwise weights for the intermediate alignments
 - * The final weight of a position is the sum of the weights of all the possible alignments supporting it
 - Maximise the pairwise alignments from the extended library with dynamic programming
 - * The score of each match corresponds to its weight
 - From the extended pairwise alignments, build a guide tree
 - Do a progressive MSA from this guide tree and the extended pairwise alignments
 - T-Coffe considers both global and local pairwise alignments and it can use information about domains and motifs

MSA benchmark

- BaliBASE was the first large-scale benchmark specifically designed for MSA
 - It is a dataset with manually refined alignments derived from structural superimposition
- BaliBASE is subdivided in several reference datasets
 - 1 - Small number of equidistant sequences
 - * This is further subdivided by identity levels
 - 2 - Families with one or more orphan sequences
 - 3 - Pair of divergent subfamilies with less than 25% reciprocal identity
 - 4 - Sequences with large extensions at the N or C terminal
 - 5 - Sequences with large internal indels
- The evaluation of the benchmark is based on a series of scores
 - The scores are evaluated only for columns that are reliably aligned in the reference (core columns)
 - Sum of pairs score (SP score): proportion of correctly aligned residue pairs in the core columns
 - Total column score (TC score): proportion of completely correctly aligned core columns
 - TC and SP score both are a number between 0 and 1
 - In a pairwise alignment SP and TC score are necessarily equal
 - In an MSA with 3 or more sequences, $SP \geq TC$
 - Both scores encourage sensitivity, but they do not test for specificity
 - * There is no penalty for wrong alignments (!)
- BaliBASE also evaluates time of execution and peak memory usage
- SP, TC, memory and time are reported as Z-scores on a spiderweb plot for each alignment tool
- What comes out of the BaliBASE benchmark?
 - No single method is perfect in all cases (!)
 - On average, consistency-based methods are more accurate but slower
 - T-Coffe suffers with N and C terminal extension
 - ClustalW and MUSCLE are the least resource-heavy tools
 - T-Coffe and MAFFT are well suited for alignments larger than those in BaliBASE
 - Multi-threading can greatly speed-up these softwares, since there is a lot of parallel computing
 - Many algos take advantage of parallel processing

Probabilistic sequence models

- A model is an object producing different outcomes (sequences) from a probability distribution
- The probability distribution in sequence space determines the specificity of the model
- The probability for model M of generating sequence s is $p(s|M)$
- In the reverse, I can see a model as an object that given an outcome computes a probability value
- Models can be trained: I can adjust the probability density function over the sequence space from a set of known sequences
 - If I want to model the globin family, I can train my model with sequences that are known to belong to that family
- After training, I can use the model to compute the probability of an unknown sequence to belong to the globin family
- The model M given a sequence s returns the probability $p(s|M)$
 - This is the probability of the model generating the sequence, not the sequence coming from the model
- Most times I am interested in the probability of a given sequence s to come from the model M
 - This is the probability of a sequence being part of a specific family
 - This is $p(M|s)$
- In order to compute $p(M|s)$ from $p(s|M)$ I need to use Bayes theorem
 - $p(M|s) = p(s|M)p(M)/p(s)$
- The priors $p(M)$ and $p(s)$ needs to be estimated to do the conversion
 - $p(M)$ is the a priori probability of any sequence belonging to the model
 - * This is the relative abundance of the class, relative to all possible classes
 - * It can be estimated from the abundance of the known sequences in the family
 - $p(s)$ is the a priori probability of the sequence and cannot be estimated reliably
- In order to avoid specifying $p(s)$ I can compare the probabilities of 2 different models
 - Instead of looking for $p(M_1|s)$, I look for $p(M_1|s)/p(M_2|s)$
 - $\frac{p(M_1|s)}{p(M_2|s)} = \frac{p(M_1|s)p(M_1)}{p(s)} \frac{p(s)}{p(M_2|s)p(M_2)} = \frac{p(M_1|s)p(M_1)}{p(M_2|s)p(M_2)}$
 - In this way, the conditional probabilities of the sequences are easy to estimate from the models themselves
 - The ratio $p(M_1)/p(M_2)$ can be estimated from the relative abundance of the 2 classes
- To make the calculation more standard, I can systematically compare any model to the NULL model
- The NULL model N is a model that generates all the possible sequences with equal probabilities, only depending on the residue frequencies

Markov Models

- HMM have their most frequent application in speech recognition
- A simple Markov Model, or Markov chain is a collection of states associated with probabilities for all the possible transitions between them
- It is useful for modeling the probability of a sequence of states that only depend on the preceding state in the sequence
- I can consider each residue as a state, and I can assume that its state depends only on the previous residue
- The Markov model will contain all the possible residues and their transition probabilities
- In this framework, the transition probability is the probability that residue B follows residue A in position x_i of a sequence
- The transition probability a_{AB} is the conditional probability of the position $i+1$ being B given that position i is A
 - $a_{A,B} = p(x_{i+1} = B | x_i = A)$
- The probability of a sequence x of length n is the product of all the transition probabilities at the various positions
 - Here I am assuming independence of each transition

- $p(x) = p(x_n|x_{n-1}) * p(x_{n-1}|x_{n-2}) * \dots * p(x_2|x_1) * p(x_1)$
- $p(x) = p(x_1) * \prod_{i=2}^n a_{x_{i-1}, x_i}$
- I can also add a BEGIN and an END state to my model for avoiding irregularities
 - * In this case the transition probability from BEGIN to a state is the probability of starting with that state
 - * This is symmetrical for transitions from a state to the END state
 - * We treat both BEGIN and END states as the same state 0 so a_{0k} and a_{j0} are transitions from BEGIN and to END
 - There is no ambiguity since transitions are only from BEGIN and only to END
- Let's say I want to model the probability that a given sequence is a CpG island
 - In such sequence, a_C, G would be much higher than elsewhere
 - I can create 2 different Markov chains M_+ and M_- for modelling the 2 sequences: CpG island and non CpG island
 - The 2 models will have the same states but different transition probabilities
 - To determine the likelihood S of a sequence x being a CpG island, I can compare the log-odds of the 2 models
 - * $S(x) = \log \frac{P(x|M_+)}{P(x|M_-)} = \sum_{i=1}^n \log \frac{a_{x_{i-1}, x_i}^+}{a_{x_{i-1}, x_i}^-}$
- I can define the probability of a sequence s_i to be generated by a family described by the model M
 - $p(s_i|M)$
- In a Markov model, the sum of probabilities going out of a state is always 1
 - It is certain that I will go out of the state
- When I have only 2 possible mutually exclusive models, I can have a measure for $p(s)$
 - $p(s|M_1) + p(s|M_2) = p(s)$
- From this, I can recover $p(M_1|s)$ and $p(M_2|s)$
 - $p(M_1|s) = \frac{p(s|M_1)p(M_1)}{p(s)} = \frac{p(s|M_1)p(M_1)}{p(s|M_1) + p(s|M_2)}$
 - $p(M_2|s) = \frac{p(s|M_2)p(M_2)}{p(s)} = \frac{p(s|M_2)p(M_2)}{p(s|M_1) + p(s|M_2)}$
- We always work with Markov models of order 1: every state depends only on the previous 1 state
 - There are also MM of order 0 or >1

Training the model

- The parameters for a model can be estimated from a set of training data
- For any sequence s and model M , I can express $p(s|M)$ as the Markov chain that can produce s
 - $p(s|M) = \prod_{j=0}^{n+1} \prod_{k=0}^{n+1} a_{jk}^{n_{jk}}$
 - In this representation 0 is the BEGIN state and $n+1$ the END state
 - The probability is the product of the transition probability for all the possible transitions to the power of how many times they do occur
- The model is always under the normalization constraint
 - $\forall j \sum_{k=1}^{n+1} a_{jk} = 1$
 - The sum of outgoing transitions from any state must sum up to 1
- Maximum likelihood estimation: the value of the parameter θ is the one that maximises the probability of the dataset D given the model and the parameter
 - $\theta_{ML} = \operatorname{argmax}_{\theta} P(D|M, \theta)$
 - The solution for any parameter θ can be obtained
 - * $\theta = a_{ik} = \frac{n_{ik}}{\sum_j n_{ij}}$
 - * The optimal value of the parameter is the frequency of occurrence of the transition in the dataset
 - * The normalization constraint forces to divide the count of transitions for the total number of outgoing transitions
- Maximum a posteriori estimation: the Bayesian correction of the ML approach
 - $\theta_{MAP} = \operatorname{argmax}_{\theta} (p(\theta|M, D))$
 - $p(\theta|M, D) = p(D|M, \theta)p(\theta)$

HMM

- Let's now try to model the presence of a CpG island in a larger sequence
 - I can integrate both models M_+ and M_- in a single model
 - I will have 2 states for each nucleotide, one for each model
 - The transition probabilities inside states of the + and - models will be similar to before
 - In addition I will have a small probability of going from a state of one model to any state of the other model
 - It will be more probable to go from - to + than vice versa
 - * This means that I will be most of the time in -, so most of the sequence is not a CpG island
- This is an Hidden Markov Model since for every position the sequence itself I cannot know which state generated it
 - For each possible nucleotide I have 2 states, and I do not know which one it came from
- Differently from Markov chains, in HMM we need to distinguish the sequence of states from the sequence of observables (symbols here)
- The sequence of states, which is hidden to us, is called the path π and it is a simple Markov chain
 - The path has transition probabilities $a_{jk} = p(\pi_i = k | \pi_{i-1} = j)$
- A symbol can correspond to multiple states but also a state can generate different symbols (!)
 - In general, the outcome of a single state derives from a probability distribution
 - We define the emission probability of symbol b from state k as $e_k(b) = p(x_i = b | \pi_i = k)$
 - The sum of emission probabilities from a state is always 1, so the state always produces something
- I can write the probability of observing the sequence x of length L under the path π
 - $p(x, \pi) = a_{0\pi_1} \prod_{i=1}^L e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}}$
 - In this equation $\pi_{L+1} = 0$, so the last transition is to the END state
 - The probability of a character x_i being generated by the model is the product of the probability deriving from the markov chain and the emission probability for that character under the current state
 - The path is typically hidden, so this equation is not useful in practice
- The probability of the above equation under a model M can be rewritten as $p(x, \pi | M)$
 - This can be decomposed as $p(x, \pi | M) = p(x | \pi, M) * p(\pi | M)$
- If I want to obtain $p(x | M)$ under an HMM I need to sum over all the possible paths
 - $p(x | M) = \sum_{\pi} p(x, \pi | M)$
- The number of possible paths is the number of states elevated to the length of the sequence
 - No way you can do that brute force
- There are different algorithms for computing $p(x | M)$ under an HMM
- Forward algorithm: I just consider the most probable path π^*