Evaluation Questions

Fernando Freire

December 12, 2018

Contents

1 Questions for evaluation **2**
1.1 Evolutionary rates. . . . 2
1.1.1 Response 1. . . . 2
1.2 Population size . . . 2
1.2.1 Response 2 . . . 2
1.3 Non-negligible mutation rate. . . . 3
1.3.1 Response 3 . . . 3
1.4 Phylogenetic trees . . . 3
1.4.1 Response 4 . . . 3
1.5 Additive distances. . . . 5
1.5.1 Response 5. . . . 5
1.6 Distance methods versus model-based methods for reconstructing phylogenetic trees. . . . 6
1.6.1 Response 6 . . . 6
1.7 AIC BIC scores. . . . 7
1.7.1 Response 7 . . . 7
1.8 Substitution rates . . . 7
1.8.1 Response 8 . . . 7
1.9 Phylogenetic networks. . . . 8
1.9.1 Response 9 . . . 8
1.10 Influence of protein stability on fitness . . . 8
1.10.1 Response 10 . . . 8
1.11 Chaperones. . . . 9
1.11.1 Response 11 . . . 9
1.12 Protein stability . . . 9
1.12.1 Response 12. . . . 9
1.13 Protein classification . . . 10
1.13.1 Response 13. . . . 10
1.14 Disordered protein . . . 10
1.14.1 Response 14. . . . 10

1 Questions for evaluation

1.1 Evolutionary rates

Explain why the ratio of non-synonymous to synonymous rates, Ka/Ks, is regarded as an indicator of positive selection. How do you interpret that Ka/Ks is significantly larger in one gene than in another one although it is smaller than one in both genes?

1.1.1 Response 1

The non synonymous substitution rate K_a is usually smaller than the synonymous rate K_s because a change in an amino acid entails more risks for the correct operation of the gene protein product and can cause the organism to lose fitness (deleterious mutation)

Why K_a/K_s is significantly larger in one gene than in another one although it is smaller than one in both genes?

This ratio has variations from gene to gene (and in sites of the same gene) depending on the influence of protein function in the organism fitness, and so tends to be lower in genes that code more fundamental proteins. In the extreme opposite case, the ratio could be larger in genes that no code protein at all (gene duplications).

Why K_a/K_s is regarded as an indicator of positive selection?

If ratio is more than one, it implies that the amino acid substitution far from causing a loss of fitness in the organisms affected, has given greater fitness advantage compared to others, and the gene is being conserved into population.

1.2 Population size

Why do we say that effective population size has the role of the inverse of an evolutionary temperature? Out of two different populations, which one do you expect to evolve faster, the smaller one of the larger one? Why?

1.2.1 Response 2

As temperature rises more microstates at a specific energy level are available to the system according to the Boltzamnn distribution of statistical mechanics, that has the form:

$$f(E) = \frac{1}{g(E/T)}$$

Where g is a increasing function depending on E/T (in fact an exponential) and f the probability that a particle has energy E (proportional to number of microstates), because if T increases, the ratio E/T decreases, g decreases and $1/g$ increases.

In detailed form:

$$f(E) = \frac{1}{exp(E/k*T)}$$

We can construct an analogy, between the probability of finding a particle of an energy level E, with the probability of finding an individual that carries a certain allele A within the population.

This probability is inversely proportional to the size of the population, since the larger the population, the more organisms with alleles of other types can be found. Let's say that the bigger the population is, the lower is the probability that an organism acquires a new allele, that is, it behaves as an inverse to the temperature with respect to energy.

Therefore, and answering the second question, small populations tend to evolve faster because they allow a greater number of alleles to coexist in the population.

1.3 Non-negligible mutation rate

Which phenomena occur when there is more than one mutation in the same genome, and there are two or more different mutants in the same population?

1.3.1 Response 3

If there is more than one mutation in the same genome, and one of the mutations is advantageous enough, the other mutations, neutral of also disadvantageous could be fixed in the population, because the overall fitness of the organism was increased.

The frequency of this mechanism depends on the location of the mutations in the genome. The closer they are, the more difficult it is for them to be separated by recombination events.

Answering the second question, these meiotic recombinations can serve to bring mutations within the chromosome, so that deleterious mutations carried by one of the progenitors are located near advantageous mutations carried by the other parent and thus pass to the offspring. Thus we would find ourselves in the previous situation, and a neutral or even non-advantageous mutation would have possibilities to be fixed in the population.

1.4 Phylogenetic trees

Consider the phylogenetic tree of n protein sequences. How many branches are there? How many pairwise distances? How many internal nodes? How many possible rooted and unrooted trees?

1.4.1 Response 4

Rooted trees We consider only binary trees. The rooted tree of n protein sequences (n leaf nodes) T_n can be defined recursively as the set formed by its internal nodes R_i and leaf nodes L_j as follows:

$$T_2 = L_2 \cup L_1 \cup R_1$$

$$T_n = L_n \cup T_{n-1} \cup R_{n-1}$$

Note: T_2 is the first case: the minimal tree has two final nodes and one internal node

From this definition we can guess some properties easily:

Number of internal nodes (cardinalR)

It seems by construction that it could be $n-1$. We will prove it by structural induction over T sets:

$cardinalR(T_2) = 1$, as T_2 only has the internal node R_1

We suppose now, that $cardinalR(T_{n-1}) = n-2$, what about T_n?

$cardinalR(T_n) = cardinalR(L_n \cup T_{n-1} \cup R_{n-1}) = cardinalR(L_n) + cardinalR(T_{n-1}) + cardinalR(R_{n-1}) = 0 + (n-2) + 1 = n-1$

So, it's demonstrated.

Number of branches (countBranches)

From each of the internal nodes start two branches. So, the number of branches is:

$$countBranches(n) = 2n - 2$$

Number of trees (countTrees)

Think now on how many ways we have to obtain recursively a tree of n nodes from a tree of $n-1$ nodes.

We can insert a new branch pairing over an existing branch. Taken into account that is irrelevant insert the branch to the right or on the left of any existing branch, we have $countBranches(n-1)$ ways to do that operation. Also we have another possibility that is generate a new branch past the root of the tree.

So we have $countBranches(n-1)+1$ possibilities of generate a n tree given a $n-1$ tree and then:
$countTrees(n) = (countBranches(n-1)+1) * countTrees(n-1) = ((2(n-1)-2)+1) * countTrees(n-1)$
so
$countTrees(n) = (2n-3) * countTrees(n-1)$, for $n > 2$
and
$countTrees(2) = 1$, trivially, because it's irrelevant the order of the two leaf nodes.
and recursively applying the formula:

$$countTrees(n) = 1 * 3 * 5 * ... * (2n-3) = (2n-3)!!$$

Number of pairwise distances
It's equal to the number of different pairs we can form with the leaf nodes.

$$\binom{n}{2} = n(n-1)/2$$

Unrooted trees We consider only unrooted trees as the derived from a star topology like the target topology used in the neighbor join algorithm, that is, n leaf nodes and one internal node. So each of the internal nodes have three branches starting from it.

The construction of such unrooted three is as follows. We take a pair of leaf nodes out from the star topology creating one internal node with three branches, one for every member of the pair of extracted nodes and other ending on the internal node of the star topology. This internal node is used as a leaf node in the remaining star topology, now a star with $n-1$ leaf nodes and the same internal node.

This process can be done recursively until finishing with a tree with one internal node, three leaf nodes and three branches, that is the minimal possible unrooted tree. The case with two branches is a rooted tree.

Number of branches
By construction, for a tree with n leaf nodes, we have
$countBranches(n) = 2 + countBranches(n-1)$
and
$countBranches(3) = 3$
$countBranches(4) = 2 + countBranches(3) = 5$
$countBranches(5) = 2 + countBranches(4) = 7$
and so on...
It seems that this should be the generic formulation:

$$countBranches(n) = 2n-3$$

Effectively, by induction over n:
$countBranches(n) = 2 + countBranches(n-1) = 2 + (2 * (n-1) - 3) = 2 + (2n-2-3) = 2n-3$
And for the first case:
$countBranches(3) = 2 * 3 - 3 = 6 - 3 = 3$
Number of internal nodes
The approach is very similar: By construction, for a tree with n leaf nodes, we have
$countNodes(n) = 1 + countNodes(n-1)$
and

$countNodes(3) = 1$
so
$countNodes(4) = 1 + countNodes(3) = 2$
$countNodes(5) = 1 + countNodes(4) = 3$
ans so on...
It seems that this should be the generic formulation:

$$countNodes(n) = n - 2$$

Effectively, by induction over n: $countNodes(n) = 1 + countNodes(n-1) = 1 + ((n-1) - 2) = 1 + (n - 3) = n - 2$
And for the first case $countNodes(3) = 3 - 2 = 1$
Number of trees (countTrees)
Think now on how many ways we have to obtain recursively a tree of n nodes from a tree of $n-1$ nodes.
We can insert a new branch pairing over an existing branch. Taken into account that is irrelevant insert the branch to the right or on the left of any existing branch, we have $countBranches(n-1)$ ways to do that operation.
So we have $countBranches(n-1)$ possibilities of generate a n tree given a $n-1$ tree and then:
$countTrees(n) = (countBranches(n-1)) * countTrees(n-1) = (2(n-1) - 3) * countTrees(n-1)$
so
$countTrees(n) = (2n-5) * countTrees(n-1)$, for $n > 3$
and
$countTrees(3) = 1$, trivially, because it's irrelevant the order of the three leaf nodes.
and recursively applying the formula:

$$countTrees(n) = 1 * 3 * 5 * ... * (2n-5) = (2n-5)!!$$

Number of pairwise distances It's equal to the number of different pairs we can form with the leaf nodes.

$$\binom{n}{2} = n(n-1)/2$$

1.5 Additive distances

Sequence distances are said to be additive when the distance between two sequences is the sum of the length of the branches that connect them. If there are n species, how many equations express the additivity conditions? Are the free parameters (branch lengths) underdetermined or overdetermined? What is the number of sequences for which the number of equations is the same as the number of parameters? For this number of sequences, which condition expresses the molecular clock hypothesis?

1.5.1 Response 5

The number of additive conditions for n species equals the number of pairwise distances $\binom{n}{2} = n(n-1)/2$.
The number of branches in rooted trees (see previous exercise), is $2n-2$.
The number of parameters (branch lengths) is underdetermined if is greater than the number of equations over distances(additive conditions):

$$2n - 2 > \frac{n(n-1)}{2}$$

For instance, with $n = 2$ we have a distance equation and two branches: $d_{12} = b_1 + b_2$ where d_{12} is the pairwise distance between the sequences at two leafs nodes L_1 and L_2, and b_i are the branch lengths.

Conversely, the branch lengths are overdetermined if the number of equations (pairwise distances) are greater than the number of parameters:

$$2n - 2 < \frac{n(n-1)}{2}$$

And it is determined if we have equality. Equality is reached for rooted trees at $n = 4$. For $n > 4$ the branch lengths are overdetermined.

If molecular clock hypothesis is true for this four sequences, all the distances from the leaves to the root must be equal.

This can be written as a set of three equations. Be $i = 1, 2, 3, 4$ the labels of the leaf nodes and 0 the label of the root node and d_{ij} the distance between nodes:

$$d_{01} = d_{02}$$
$$d_{01} = d_{03}$$
$$d_{02} = d_{03}$$

Because the six branch lengths are determined uniquely by the six equations, there is nothing we can do to force ultrametricity. But molecular clock conditions also could be satisfied depending on the values of the distances, i.e, the three additional conditions being linearly proportional to the other six equations.

1.6 Distance methods versus model-based methods for reconstructing phylogenetic trees

Which are the advantages and disadvantages of these two types of methods?

1.6.1 Response 6

Advantages Model-based methods vs distance methods:

1) Don't use distances as the basis of branch length calculations. For this reason are more accurate than distance method. The use of distances has a lot of well know problems:

- Construction of the tree heuristics. Main methods do strong biological assumptions as additive distances (NJ) or ultrametricity (UPGMA). Also the trees are prone to errors derived for the errors in pairwise alignments, that in CLUSTAL remains unchanged all over the computation. More sofisticated methods (T-COFFEE or MUSCLE), try to avoid this problems, allow some changes in the initial alignments (T-COFFEE) and in the tree topology, that is explored iteratively (MUSCLE).

2) Model methods are based on a consistent and compact model of evolution, and have a more solid biological basis. Distance methods are a eclectic mixture of biological concepts.

Advantages Distance method vs Model-based methods

1) Less computational time. Model based methods have NP complexity. Distance method give a relatively fast tree and somehow accurate.

2) ML works on samples from tree population, because it's impossible to explore the whole population of trees: this population is huge and grows O(n!) with the number of sequences (see exercise 4). There are local optimization methods but is not guarantee that the tree selected was not in a local maximum.

Both types of methods share also problems related to the alignment strategies:

- Correct selection of true orthologs as leaf nodes.
- Dependence of alignment method: parametrization and scoring (substitution matrices, gap penalties, ...). A subtle difference in the choice of parameters could originate a different tree. In particular the scoring of gaps, gap open penalty and gap extend are crucial and with no clear biological basis.
- Alignment method as is. Heuristics. Multialignment is NP complete and is not possible to do a n-dimensional "multiwise" precise alignment. Commonly the starting point is the whole set of pairwise comparisons between every pair of sequences. The pairwise alignment has to be done by heuristic approaches also, if the number of sequences is huge.

1.7 AIC BIC scores

Why are the AIC or BIC scores needed in order to test alternative models of molecular evolution?

1.7.1 Response 7

Models with more parameters yield higher likelihood than models with less parameters (overfiting)

So, it's necessary a strategy to score the testing tree penalizing the number of parameters.

AIC and BIC are statistical estimators of the relative quality of the models. The more generic and also simpler quality control formula is such this:

$$q(p, L) = k_1 * q_1(p) - k_2 * q_2(L)$$

where p is the number of parameters, L the likelihood of the tree, k_1 and k_2 two positive constants and q_1 and q_2 two increasing funcions. The model with lower Q is the preferred.

AIC and BIC are in accordance with this general formula.

1.8 Substitution rates

Which features determine the among-site variation of the substitution rate? How is this variation implemented in empirical models of the substitution process?

1.8.1 Response 8

The functional of the gene or protein determine the among-site variation of the substitution rate:

1. Proteins performing fundamentally functions tend to evolve more slowly than other proteins.
2. Some protein domains need to be evolutionary conserved more than other domains of the same protein because they participate more strongly on the protein function (active sites). In that domains, substitution rate is lower than other domains.
3. At gene level, third and second codon positions are less conserved than the first codon position, because the first position determines more strongly the amino acid at translation step than the second position, and the second codon position more strongly than third codon position, so substitution rates (r), verify $r_1 < r_2 < r_3$.

Empirical models

They are based on the hypothesis that each site has a rate that is determined by its position in the molecule, regardless the residue that occupies the site.

We can typify two types of approaches.

1. Step wise. Dividing the gene/protein in two or more regions of different substitution rates. The simplest model take into account invariant sites, at which the sequence is considered fully conserved.

2. Continuous wise. Gamma distributions. The transition from active sites with low or null substitution rate to sites with high rate are modeled with a function depending on a shape parameter. The larger the shape parameter α, the smaller the substitution rate.

1.9 Phylogenetic networks

When is it suitable to model evolutionary relationships with a network rather than with a tree, despite the increase of complication?

1.9.1 Response 9

Generally the genetic information flows not only in the direction from one evolutionary parent. If we consider that these alternative gene flows or alternative genetic sources are relevant, we need to improve the topology of genetic relationship to a more flexible model that take into account all possible paths to exchange genetic information. In this case we need to adopt a network topology, a hierarchical binary tree is not enough.

Alternative genetic information paths are:

1. Recombination between parents. If a species has two parents, both parent are evolutionary related to it, but they could be unrelated among them.

2. Hybridation.

3. Horizontal gene transfer.

4. Incorporation of genomes (Margulis hypothesis).

1.10 Influence of protein stability on fitness

How do we model the influence of protein stability on fitness? Explain why we say that the evolution of protein stability is largely a neutral process, and in which range of parameters this is true.

1.10.1 Response 10

The fitness are modeled as the probability to find the protein in the native state, in function of free Gibbs energy of the transition folded state. The more negative ΔG_{folded} the larger the fitness, because the transition to folder state is more likely to occur spontaneously, i.e., the reaction $protein_{unfolded} \rightleftharpoons protein_{folded}$ is more biased from left to right.

$$\Delta fitness = fitness_{folded} - fitness_{unfolded}$$

$$\Delta fitness \cong \frac{1}{1 + e^{\Delta G_{folded}/k_B T}}$$

This way of thinking assumes that the protein develops all its functionality at a folded native state, and thus this the more folded versions of the protein, the the best fitness to organism. But it's not always true, there are natively unstructured proteins and the protein function could depend on the linkage to another proteins or compounds of the system.

There are a lot of experimental studies showing that the mutations often have no relevant effect over protein's structure or function. It seems that protein evolution can described by Kimura's neutral theory of evolution: most genetic change is due to the stochastic fixation of neutral mutations, that not change the organism fitness.

From the previous equation, if temperature is low or the protein is enough large $|\Delta G|$ high, that is $|\Delta G| >> KT$ denominator goes to a very high value, and the variation in fitness tend to zero.

As the effect of a mutation on stability is $\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild-type}$, we have

$\Delta G_{mutant} = \Delta G_{wild-type} + \Delta\Delta G$

So the limit condition that requires $|\Delta G| = |\Delta G_{mutant}|$ be large, imposes also that $|\Delta G_{wild-type}| >> \Delta\Delta G$

So the more stable the wild type of the protein is, more neutrality is reached, and this favors evolvability.

1.11 Chaperones.

Describe the role of chaperones in evolution

1.11.1 Response 11

Chaperones help the folding of unfolded and wrongly folded proteins, preventing aggregation.

So chaperones confer to the system: protein, substrate and the same chaperones a $|\Delta G_{folded}|$ higher than they would have if they were not easing the folding.

Therefore confer more stability to the protein and then, according to what was discussed at the end of the previous exercise, favor their evolution, increasing their ability to store non-deleterious changes, or what is the same, accumulating possible new functions.

Also, chaperones buffer deleterious mutations, by increasing their expression levels.

1.12 Protein stability

Describe the trade-offs and synergies between stability against unfolding and stability against misfolding.

1.12.1 Response 12

A protein primary structure needs, to achieve proper functioning, to follow two structural pathways to the 3D structure:

1. To ease the folding to a stable three-dimensional shape. Then to have a suitable unfolding stability.
2. To provide appropriate residues for binding with other proteins or ligands. So to have a adequate misfolding stability.

This originates a trade-off: the protein folding process needs to deal with two constraints at the same time: that the fold is stable but not stable enough to block the functional residues needed to the proper protein function.

I think on two types of synergies to face these difficulties:

1. Temporal synergies, based on the thermodynamic oscillation between less compact (misfolded)and more compact (folded) states, so that the necessary interactions can be favored in short time.

2. Spacial synergies. Only those residues whose properties do not contribute to function must be the basis for the folding, and the segments containing functional residues needed for binding must to be the hardest to fold.

1.13 Protein classification

How do protein proteins structures change when their sequences diverge? Is it justified to assume that homologous proteins (same superfamily) always belongs to the same structural class (fold)?

1.13.1 Response 13

The RMSD of protein sequences increases exponentially with identity percentage so that the structures are very tolerant to changes in sequences, i. e. structure divergence is evolutionary slower than sequence divergence: selective pressure is stronger over protein structure than over protein sequence, structure is strongly conserved by evolution.

This scenario encourages us to classify proteins into families based on the homology between their sequences. They can be aligned following one of the known algorithms of alignment.

However, as we commented about exercise 6, there is some aspects of those algorithms that reside in the correct score of the gaps, with little or no biological basis. These gaps can separate or close areas of the proteins that interact with each other during the bending process, significantly modifying their native state.

Therefore, a good alignment, with a high percentage of identity, but that includes sequences of gaps in relevant sites, may be overestimating the three-dimensional analogy of the proteins that are being compared.

Another aspect of the alignment of sequences that can be misleading, is that they do not take into account possible internal coupling (that could be even co-evolutionary), between certain residues that interact to achieve the native conformation. These residues should be anchor characters with a higher algorithmic score to favor their 1-D alignment.

1.14 Disordered protein

Which are the evolutionary properties of disordered proteins that differentiate them most from structured proteins?

1.14.1 Response 14

Intrinsically Disordered Proteins (IDP) lack a unique 3D structure in native conditions. They maintain a huge number of different conformations in dynamic equilibrium. This gives them a great flexibility to adopt various conformations which is why they usually perform regulatory functions(for example, anchoring to DNA regulatory or affinity to be coupled to co-activators in signaling pathways).

Protein structure is generally more conserved than sequence as we stated in previous exercise, but is it true for regions or entire proteins that can adopt different structures in different situations?

One can expect that these proteins are more prone to accumulate neutral changes than structured proteins, precisely because of their conformational flexibility that would help them to maintain their function in the face of small changes. Therefore, it is expected that they will evolve more quickly. They may also be able to absorb larger changes. In fact it seems that in the evolution of these proteins, large indels are produced much more frequently than in structured proteins.