# Protein Evolution

**Problem presented by**

Trever Greenhough

*Centre for Molecular Biomedicine, Keele University*

**Problem statement**

Among the proteins that have evolved over hundreds of millions of years, with important roles in defence against invading micro-organisms, are the pentraxins. The two major members of the family are known as CRP and SAP; and they evolve due to mutations in the underlying DNA. The Study Group was asked to construct a model of this evolution in order to answer specific questions about the occurrences of these proteins in man and in the horseshoe crab.

**Study Group contributors**

C. J. Chapman (Keele University)
J. Gravesen (Technical University of Denmark)
P. G. Hjorth (Technical University of Denmark)
R. Ketzscher (Cranfield University)
A. A. Lacey (Heriot-Watt University)
G. Richardson (University of Nottingham)
N. D. Stringfellow (Cranfield University)

**Report prepared by**

P. G. Hjorth

# 1 Introduction

Many of the proteins found in modern man and other successful[1] species have evolved over hundreds of millions of years through various lineages and species. Many of these proteins have important roles in defence against invading micro-organisms and thus the evolutionary changes have important restrictions in order to maintain function albeit in a global rather than a specific sense.

One such family of proteins is the *pentraxins* which play a key role in innate (non-adaptive as opposed to antibody based) immunity, the two major members of the family being C-reactive protein (**CRP**) and serum amyloid P-component (**SAP**).

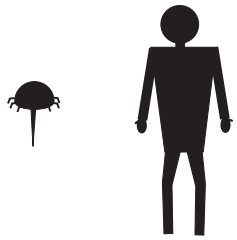

Figure 1: Horseshoe Crab (*Limulus polyphemus*), Man (*Homo Sapiens*)

Both proteins have been found in all species in which they have been sought, and in particular in the ancient invertebrate 'living fossil' *Limulus polyphemus* (the horseshoe 'crab') and in *Homo Sapiens* (man), where CRP is the major acute phase reactant produced in response to tissue damage and inflammation. CRP levels are routinely and universally measured in man as a clinical indicator of underlying infection.

# 2 Proteins

Proteins are composed of long chains of amino acids (around 200 amino acids in the case of the pentraxins); there are 20 different kinds of amino acids. Some of the amino acids in a protein are important in maintaining the structure of the protein and some are essential for its function. These requirements will vary depending on the protein involved. Each amino acid is coded for (defined by, produced by) a *triplet* of nucleotide bases (a codon) in the relevant piece of DNA. Rather than coding for an amino acid, some DNA triplets are 'Stop' codons signalling the end of protein synthesis. There are four different bases: C, G, A and T. A single change in one of these bases, from one of the four bases to another, may produce a change in the resulting amino acid and hence the final protein.

**Note:** Since there are only 20 amino acids, the genetic code is degenerate with some amino acids specified by sets of codons. A change in the third base of the codon often results in no change of the amino acid, while a change in the first or second base of the codon usually does.

[1] In the strictly evolutionary sense of being still around at present.

# 3 Mutation and Evolution

A suitable simple view of protein evolution is that it arises from random changes (mutations) in amino acids, resulting from random changes (mutations) in the relevant coding DNA, and that these mutations are dominated by *point mutations* where a change in a single DNA base occurs. Other DNA mutations, not to be considered here, include 'insertions', 'deletions' and 'frameshifts'.

Point mutations may have a variety of effects. Of these we consider here only the 'acceptable' (benign or beneficial) mutations; those that are passed on to further generations.

A commonly used model is that **P**oint (**A**cceptable) DNA **M**utations (PAMs) occur at a constant rate $\mu$ of approximately 20 PAMs/100 million years = 200 PAMs/Gyr (1 Gyr = $10^9$ years).

In this model, two proteins of 200 amino acids (coded by 600 bases) which have diverged over 500 million years (a total of 1 Gyr of evolutionary divergence) would show a difference of 200 bases and the two genes would be 400/600% *i.e.* 67% homologous (identical). Translating this in to amino acid homology (which is in practice considerably less than the DNA homology) is not straightforward; in the extreme cases 200 base changes could produce 200 amino acid changes (0% amino acid homology) or they could produce none (100% homology). The position and nature of the mutation are clearly of paramount importance, but account needs to be taken of the degeneracy in the genetic code, the possibility of mutation back to a previous state, and the restrictions imposed by preservation of structure and function (exclude detrimental mutations). Basing the PAM model on the mutation rate given above ($\mu$ = 1 PAM/5 million years) fails spectacularly when the new data [1] given below is considered.

# 4 The Pentraxins

In simple terms, some 500 million years ago (at least) an evolutionary divergence in coelomata, the ancestor(s) of both chordates (eventually leading to vertebrates and humans) and arthropods (leading to, for example, spiders and the horseshoe crab) led to the establishment of two new evolutionary lines. Note that present day man and the horseshoe crab represent over 1,000 million years ($2 \times 500$) of evolutionary divergence.

It has recently been shown [1] that the two proteins CRP and SAP are both present in the horseshoe crab, while others have shown the presence of both in man and in every other species in which they have been sought. This is consistent with the view that the two proteins arose from a common ancestor protein (say CRP) via a gene duplication event (creation of a new, additional gene by duplication of an existing gene) and that they have been evolving and diverging since. It is not consistent with the accepted view, arising from using $\mu$=1 PAM/5 million years = 200 PAM/Gyr in the PAM model, and from the previous assumption that SAP did not exist in invertebrates, that the duplication event occurred hundreds of millions of years after the divergence of the two evolutionary lines which lead to man and the horseshoe crab. The two proteins show 51% amino acid homology in man (and a similar value in other mammals) and 34% in the horseshoe crab.

# 5 Parameters and assumptions

A1 The protein SAP was first generated from CRP by gene duplication say 500 million years ago, at which point (time zero) the two genes and proteins (approximately 600 bases and 200 amino acids) were identical.

A2 If we need to define the initial amino acid sequence in terms of relative abundance of the 20 amino acids, this can be reasonably based on the numbers of coding triplets (e.g., 6 times more of the amino acid Arg than Trp).

A3 The proteins SAP and CRP have been diverging through random point acceptable mutations of the coding DNA for 500 million years and are now 51% homologous in man and 34% in the horseshoe crab at the amino acid level. The rate $\mu$ of random point acceptable mutations is constant (current thinking suggest $\mu$ has the same value for all species, around one base every 5 million years).

A4 A certain percentage of the original amino acid sequence of CRP, and hence of the evolved, present day CRPs and SAPs, is required to remain constant by the constraints of structure and function. A reasonable estimate in terms of amino acids is 20% in man (the same 20% in both proteins) and 10% (again the same 10%) in the horseshoe crab.

# 6 The problems posed to the Study Group:

Q1 In (a) humans (b) the horseshoe crab, what is the future steady-state minimum amino acid homology (in percent) between the two proteins SAP and CRP (this will be independent of all parameters except the constant percentage of amino acids?).

Q2 Can we now deduce $\mu$ for both species?

Q3 Can we now determine when (from time zero) the minimum homology will occur?

Q4 Can we deduce the homology between the SAME protein (CRP) in man and the horseshoe crab?

# 7 A simple model of base evolution due to point mutation

To model the probabilistic evolution of a single base (say, A), we define a state 'vector' $P \equiv (A, G, C, T)^{\perp}$, whose short ($\delta t$) time transition matrix is given by

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1-3\mu\delta t & \mu\delta t & \mu\delta t & \mu\delta t \\ \mu\delta t & 1-3\mu\delta t & \mu\delta t & \mu\delta t \\ \mu\delta t & \mu\delta t & 1-3\mu\delta t & \mu\delta t \\ \mu\delta t & \mu\delta t & \mu\delta t & 1-3\mu\delta t \end{pmatrix} \end{array}.$$

The associated $\mathbf{Q}$ (transfer) matrix is given by

$$\mathbf{Q} = \begin{bmatrix} -3\mu & \mu & \mu & \mu \\ \mu & -3\mu & \mu & \mu \\ \mu & \mu & -3\mu & \mu \\ \mu & \mu & \mu & -3\mu \end{bmatrix}$$

which has eigenvalues 0 and $(-4\mu)^3$. Thus the fundamental solution is given by

$$\mathbf{P}(t) = \exp(\mathbf{Q}t).$$

The probability of $A$ occupancy for a single site at time $t$, known to be in state $A$ at time 0 is then

$$P_A(t) = \frac{1}{4}(1 + 3e^{-4\mu t})$$

and for the probability of later occupancy by any other non-$A$ nucleotide (*e.g.* $C$), is

$$P_C(t) = \frac{1}{4}(1 - e^{-4\mu t}).$$

Note that $P_A$ is a decreasing, and $P_C$ an increasing function as they should be. Note also that the (nonlinear) timescale for change is about $1/4\mu$.

# 8 Evolution of amino acids

Suppose we say that an amino acid is specified by two bases (see the remark in the section on amino acids), then a state which is initially $AA$ gives rise to the following probabilities (the notation should be self-evident):

$$\begin{aligned} P_{AA}(t) &= \frac{1}{16}\left(1 + 3e^{-4\mu t}\right)^2 \\ P_{AC}(t) &= \frac{1}{16}\left(1 + 3e^{-4\mu t}\right)\left(1 - e^{-4\mu t}\right) \qquad \text{(6 different of these)} \\ P_{CC}(t) &= \frac{1}{16}\left(1 - e^{-4\mu t}\right)^2. \qquad \text{(9 different of these)} \end{aligned}$$

Suppose we are faced with a string of independent base pairs (*i.e.* forming a chain of amino acids). Suppose further that each amino acid has evolved from some initial state, displaying now, for man, at time $t$ a 51% homology between CRP and SAP, and that this evolution has taken place under the constraint of keeping $20\% = 1/5$ fixed, then the combination $\mu t$ must have a value such that

$$\frac{1}{5} + \frac{4}{5}\left(P_{AA}^2 + 6P_{AC}^2 + 9P_{CC}^2\right) = \frac{51}{100}. \tag{1}$$

Putting $t = 0.5$ Gyr, we can solve this equation for $\mu$. This gives

$$\mu \approx 0.17 \text{ Gyr}^{-1}.$$

A similar calculation for the crab (keeping $10\% = 1/10$ fixed) gives

$$\mu \approx 0.26 \text{ Gyr}^{-1}.$$

Now we can begin to answer the questions posed (see section 6).

Question: What is the steady-state (long-time, minimum homology) in this model?
Answer: Take the limit $t \to \infty$ for the right-hand side of equation (1). For man 25%, for crab 16%.

Question: What is the mutation rate?
Answer: $\mu = 0.17 - 0.26$ Gyr$^{-1}$ (technically, this is the time scale for the approach to the steady state). Note: this calculation implies different mutation rates for man and crab, with the crab having about 1.5 times higher mutation rate.

Question: What is the time to reach equilibrium?
Answer: $1/(4\mu) \approx 1 - 1.5$ Gyr.

Question: Can we deduce the homology between corresponding proteins in crab and man?
Answer: Yes. The above numbers are consistent (with a small relative difference in the mutation rates) if the SAP and CRP proteins have evolved away from each other for the entire 0.5 Gyr period. Thus, in the figure below, the protein branching should coincide with the species branching.

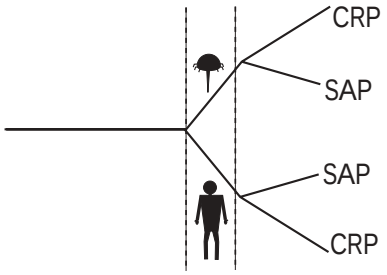


Figure 2: One model for the evolutional branching tree.

# 9 Like proteins in crab and man

In this section we consider the homology for a given protein (CRP or SAP) between crab and man. Specifically, we assume that the distinction between the 10% 'fixed' and the 20% 'fixed' amount of amino acid in crab and man respectively is itself a result of evolution. Introducing $f_c$ and $f_s$ (for crab and (homo)Sapiens) fractions of 'fixed'

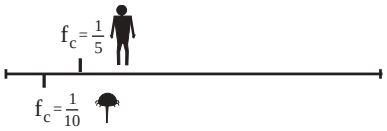


Figure 3: The 'fixed' part of the proteins in man and crab

protein, the evolution equation for the homology $H$ as a function of time takes the form

$$H(t) = f_c + (f_s - f_c)P_{AA}(t) + (1 - f_s)(P_{AA}(t)^2 + 6P_{AC}(t)^2 + 9P_{CC}(t)^2).$$

We have assumed here that the fixed part of the horseshoe crab's DNA is part of that in man's. Note that $H(0) = 1$, as it should. Furthermore, note that the value $H_\infty = f_c + \frac{1}{16}(f_s - f_c) + \frac{1}{16}(1 - f_s) = \frac{1}{16}(1 + 15f_c) \approx 15\%$ predicts the asymptotic (final) homology between either protein in man and either in the horseshoe crab.

Using $t = 0.5$ Gyr, $f_s = \frac{1}{5}$, and $f_c = \frac{1}{10}$, and the extra information that $H = 25\%$, we can solve for $\mu$ and find that $\mu \approx 0.43$ Gyr$^{-1}$.

Suppose that we now use this value of $\mu$ for both species, and again turn to the evolution equation of section 8. Then we can solve (each species equation) for the evolution time $t$.

The result is that in homo sapiens (with SAP/CRP ratio 51%) the protein divergence time of about 0.20 Gyr, whereas in crab (with $SAP/CRP$ ratio of 34%), the protein divergence time of about 0.31 Gyr.

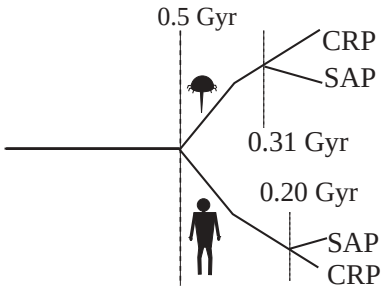


Figure 4: Based on slightly different assumptions, an alternate model for the evolutional branching tree.

This model, which ties the evolution rate to the same value in both species, results thus in a prediction of separate divergence times for (SAP/CRP). This idea is illustrated above (and is contrary to the assumption A3 of Section 5).

# 10 Conclusion

We have set up (in section 8) the simplest possible mathematical model for random point mutation, and used the model to answer the question posed to the group. We have also shown (in section 9) how one can fairly easily vary the assumptions to arrive at other conclusions. The model itself obviously can (and should) be refined to take into account a number of biological facts as well as the various types of mutations that occur in real life.

# References

[1] A. K. Shrive, A. M. Metcalfe, J. R. Cartwright and T. J. Greenhough, *C-reactive proteins and SAP-like pentraxin are both present in Limulus polyphemus haemolymph: Crystal structure of Limulus SAP.* Journal of Molecular Biology, **290**, 997-1008 (1999).

[2] T. H. Jukes, C. R. Cantor, Evolution of Protein Molecules pp. 21-123 in: *Mammalian Protein Metabolism*, H. N. Munro (Ed), Academic Press, New York (1969).

[3] W-H. Li, *Molecular Evolution*, Sinauer Associates, ISBN 0-87893-463-4, Sunderland MA (1997).

[4] L. W. Nichol and D. J. Winzor, Chapter 9 in C. Frieden and L. W. Nichol (Eds), *Protein-Protein interactions*, Johns Hopkins University Press, 1983.

[5] E. Zuckerkandl and L. Pauling, *Evolutionary divergence and convergence in proteins*, pp 97-166 in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel (Eds), Academic Press, 1965.