Pacific Symposium on Biocomputing 5:66-77 (2000)

THE EVOLUTION OF DUPLICATED GENES CONSIDERING PROTEIN STABILITY CONSTRAINTS

D.M. TAVERNA*, R.M. GOLDSTEIN*†
**Biophysics Research Division, †Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055, USA*

We model the evolution of duplicated genes by assuming that the gene's protein message, if transcribed and translated, must form a stable, folded structure. We observe the change in protein structure over time in an evolving population of lattice model proteins. We find that selection of stable proteins conserves the original structure if the structure is highly designable, that is, if a large fraction of all foldable sequences form that structure. This effect implies the relative number of pseudogenes can be less than previously predicted with neutral evolution models. The data also suggests a reason for lower than expected ratios of non-synonymous to synonymous substitutions in pseudogenes.

1 Introduction

Gene duplications are quite common among present-day organisms[1]. Duplication events can range anywhere from reverse transcription of mRNA back into the genome, transmission of vectors between members of a population, a tandem replication caused by an offset crossover event during mitosis, or the copying of an entire chromosome or genome[2]. It has been speculated[3,4,5,6,7] that duplication events are more capable of producing novel gene function than taking a single existing protein coding region and point mutating until a new function is found. The successfulness of such strategies may be evidenced by the frequent discovery of new intra-genome multigene families in both Prokaryotes and Eukaryotes. These families consist of a set of many homologous functional genes that exhibit a high degree of gene redundancy, meaning they are often quite similar in structure, sequence, and functionality[3,8]. Alternatively, duplicated genes can become pseudogenes, sequences of DNA clearly related by sequence similarity yet whose message is not translated to product.

It is thought that unprocessed pseudogenes form when a duplicated gene evolves with no selection pressure for protein activity. However, the ratio of functional genes to pseudogenes within gene families is considered too high given the rate of deleterious mutations to be explained by neutral evolution. Additionally, the ratio of non-synonymous to synonymous substitution rates in pseudogenes are lower than one would expect at random. This leads one to suspect that there are some conservation mechanisms selecting for amino acid sequence similarity. Selectionist theories explain this by imposing functionality arguments that bias advantageous mutations. However, we have no idea what the actual selection pressures are or what the magnitude of these pressures might be in such a situation.

It is possible to rationalize functional gene redundancy and the present day ratio of functional genes to pseudogenes using ideas from neutral theory, without having to impose selective constraints on functionality. We propose that simply the constraint of having active genes produce foldable proteins is enough to produce the previously mentioned behavior. Specifically we explore how the pressure to fold influences the structure of the duplicated gene's protein product. We present a haploid population evolution model of gene duplication in which we observe the change in protein structure over time.

1.1 Previous Models

Proteins have three major evolutionary constraints. They must fold to a structure in a reasonable time, the structure they fold to must perform a task, and the folded structure must be stable enough to perform that task reliably. The methods they use to conform to such constraints have been a great source of debate for the past few decades[9]. The two main views of evolution, and also therefore gene duplication, are neutralist (most mutations are either deleterious or neutral) and selectionist (most mutations are either deleterious or advantageous). The following is a short review of some of the models of the past involving these theories.

Neutralist models: A common interpretation of the processes that affect duplicated genes is that one of the genes is selected to keep its original function while the other is free to mutate and search sequence space without any selective constraint[2,3,6]. The main problem of this first theory is that many more negative than positive mutations exist for functional proteins. As such, it is expected that without selective constraints a pseudogene will form quite quickly. But recent studies have found that the majority of gene families have a high percentage of functional members, more than one would expect given the probability of lethal mutations[2,3]. This indicates that the coevolution of duplicated genes is quite common.

Selectionist/Functionalist models: This second category focuses on the effect the gene has on the organism. It is common test among models to monitor the selection pressure that balances the effect of the deleterious mutations[3,8,10]. While these models explore the effect of mutations on balancing selection, the effect of phenotype changing due to mutation has yet to be explored in detail. When one considers the space of all possible sequences connected via point mutations, we find that correlations exist between neighboring sequences and the structural phenotypes of the sequences. Therefore, the primary sequence of a protein determines its structural robustness to mutation. It has been shown using lattice models that as the pressure to fold increases, walks on the sequence landscape become glassy and localized to a particular structure[11]. For realistic values of folding pressure, where only a small fraction of possible sequences is foldable, the rate of deleterious

mutations is highly dependent upon the percent of sequences that fold into a specific structure *k*, also known as the designability ν_k[12,13].

Recently, other models have been suggested that assume genes have multiple functions. The preservation of both genes is due to a partitioning of original function between the two new genes[8,14]. The most recent of these is the complementary degenerative mutation (subfunctionalization) model. This is an exciting theory which incorporates the entropy of establishing shared functional regulation networks between the duplicated genes. It is also one of the first real attempts to explain what happens between the duplication event and pseudogene formation by monitoring the gene's regulation and activity. Still there are some possible criticisms of this model. For instance, the genome's reason for keeping the duplicated protein is still selection pressure for total activity. Also, the original functionality can become diluted with time and sensitivity to mutations increases as the function is spread through the networked gene family. Since genes found in nature have regulatory elements near to the original gene, it is possible such regulatory networking that result from this model may be selected against. Finally, this model selects for genes with larger numbers of independent regulatory units.

1.2 Our Approach

A factor lacking in all these models is the process by which a duplicated gene becomes a pseudogene. Between these two extremes, the phenotype of the translated protein is free to change. Deleterious mutations most often affect the gene's protein product and rarely the transcription process. Yet it is most often a mutation in the regulation region that leads to a lack of gene expression. Assuming this is relatively constant for most genes, this effectively sets a window of opportunity in which the gene must search phenotype space for a needed function before the gene ceases to be expressed. We would like to create a model that allows for the searching of sequence space without requiring a selection for function (since we have no idea what the actual selection pressures might be), yet still allows the molecule to retain the statistical possibility of regaining function should the need for a novel function arise. We have developed a basic model of protein folding that allows us to capture some real-life characteristics of mutation and phenotype expression. We further construct an evolutionary model where we assume a gene duplication has occurred and spread throughout the population. We use these models to address three issues central to the understanding of gene duplication effects. First, we apply our model to explain why a seemingly disproportionate number of gene duplications might be allowed to evolve to functional genes. Second, we investigate how a structure's designability influences the structure and sequence similarity often found in multigene families. Finally, we suggest a reason for why the non-functional pseudogenes might exhibit a

ratio of non-synonymous to synonymous substitution rates that is lower than one would expect due to pure neutral evolution.

We find that for the cases we studied, sequences resulting from a gene duplication event that adopt highly designable native structures are more likely to preserve their structure than lower designable structures during of population evolution. Specific ramifications of this effect will be discussed in the conclusion including testable scenarios.

2. Methods

We present a simplified model of both protein folding and the process of evolution. We find that it is with this approach that we can best deal with the complex biological issues associated with these processes, and yet still maintain computational tractability.

2.1 The Model Protein

Our protein model consists of a chain of 25 monomers, confined to a 5x5 two-dimensional maximally-compact square lattice, with each monomer located at one lattice point. This provides us with 1081 possible conformations represented by the 1081 self-avoiding walks on this lattice. We do not include structures related by rotation, reflection, or inversion in our analysis. We assume that the energies of any sequence in conformation *k* is given by a simple contact energy of the form:

$$\sum_{i<j} \gamma(A_i A_j)\Delta_{ij}^{k} \qquad (1)$$

Here, Δ^{k}_{ij} is equal to 1 if residues *i* and *j* are not covalently connected but are on adjacent lattice sites in conformation *k*, and $\gamma(A_iA_j)$ is the contact energy between amino acids A_i at location *i* and A_j at location *j* in the sequence. These contact energies represent those derived by Miyazawa and Jernigan based on a statistical analysis of the database of known proteins, and implicitly includes the effect of interactions of the protein with the solvent[15]. There are 132 pairs of residues that can possibly come into contact, with 16 of these contacts present in any given structure.

2.2 The Folding Model

A characteristic universal to most all proteins is the ability to be stable in their folded state. Reasons for stability include functionality, avoiding proteolysis,

aggregation, or initiating an immune response. Several diseases are known to exist due to the presence of misfolded protein states (including prion and sickle-cell anemia diseases). Although gene concentration and rate of protein misfolding are other important parameters that influence the evolution of an affected population, they will not be considered here in this simplified model. We characterize a protein's stability with its ΔG of folding, defined as

$$\Delta G = -kT \ln\left(\frac{P_f}{P_u}\right) \tag{2}$$

where P_f and P_u refer to the probability finding a given sequence folded in its native state and unfolded state, respectively. If Z is the partition function,

$$Z = \sum_i e^{-E_i / kT} \tag{3}$$

then

$$P_f = \frac{e^{-E_f / kT}}{Z} \tag{4}$$

Using

$$P_f + P_u = 1 \tag{5}$$

we get

$$\Delta G = E_f + kT \ln\left(Z - e^{-E_f / kT}\right) \tag{6}$$

We make the assumption that the thermodynamic hypothesis is obeyed, and that the lowest energy structure is the native state[16]. We set kT equal to 0.6 as is the value to be used with the Miazawa-Jernigan potentials, corresponding to a temperature of 298 K[15].

The lower bound for the structural stability is such that if ΔG is greater than ΔG_{crit} then the protein sequence is considered unfolded. This requirement for stability is biologically justified as it makes the bulk of the sequences unstable while

preserving a sufficient amount of stable sequences for adequate mutational sampling. We do not consider activity here although it is possible that functionality will eventually be found under the constraint of adequate stability.

2.3 The Evolution Model

As natural proteins are the products of population evolution, our models incorporated this idea as a way to explore the fitness landscape in a biologically relevant manner. We chose to measure the structural evolution of gene products resulting from the duplication of genes originally coding for one of two structures; one from among the 10% most designable structures and one from among the 10% least designable structures. Structural similarity as determined by <*Q*> is:

$$\langle Q \rangle = \left(\frac{1}{16 Npop^{f}} \right) \sum_{n=1}^{Npop^{f}} \sum_{i<j} \Delta_{ij}^{I} \Delta_{ij}^{n} \qquad (7)$$

Here, the original structural contacts are defined by Δ^{I}_{ij}, and Δ^{n}_{ij} represents the contacts for the nth member of the total population of foldable proteins $Npop^{f}$. We performed our evolution on a population of 1000 protein sequences. Eight trials were run for each structure using randomly chosen viable sequences with initial $\Delta G < -2.0$ and $\Delta G_{crit} = 0.0$ Over the entire population the probability of each amino acid mutating was 0.0004 per residue position. In practice this was achieved by choosing the total number of point mutations in the population from a Poisson distribution with average 10. Both the population size and mutation rate were chosen to be comparable to previous analytical models of evolution processes[1,17,18]. The mutations were distributed randomly across all residue positions in the population. We considered this analogous to the way mutations are introduced during a reproduction step. The stability of each protein in the population was then calculated using equation 6.

To determine the members of the next generation, we used a tournament selection method that biased reproduction towards those sequences with $\Delta G < \Delta G_{crit}$. This served as a way to tune the penalty for being unfoldable. This was incorporated because sometimes natively unfolded proteins only weakly affect survival rates[5,19]. Ω members were randomly drawn, with replacement, from the population. The surviving sequence was chosen randomly from among all sequences with stabilities of $\Delta G < \Delta G_{crit}$. If no stable sequences were drawn, the surviving sequence was chosen randomly from among all unfoldable sequences. Only one sequence was retained as the result of each comparison. Ω is therefore defined as the number of sequences that must compete for the offspring to survive to the next generation. Note

that when $\Omega = 1$ we have completely neutral evolution. This process was executed 1000 times for the purpose of maintaining a constant population.

To make the model more stochastic, Ω was chosen from a modified Poisson distribution. When selecting a number from a Poisson process with average Λ, there is probability $e^{-\Lambda}$ that you will choose zero sequences from the population to compare. In practice we selected the number of sequences to be drawn by using a Poisson process of parameter ω-1 and added 1 to the result. This gave us Poisson statistics offset to guarantee selection of at least one sequence during each tournament. For this paper, we set $\omega = 1.5$ which allowed for about 0.3% of the population to remain unfoldable in equilibrium when $\Delta G_{crit} = 0.0$.

Each population dynamics trial was allowed to run for 1,000,000 generations. During this time, the foldable population's average structural similarity to the original sequence's structure was calculated. Additionally, we also calculated the average designability adopted by the foldable portion of the population, defined as the proportion of all random viable sequences that formed into that structure. Both of these quantities were averaged over all 8 trials.

3. Results

3.1 Low values of ΔG_{crit} makes sequence connectivity analogous to biological situations.

We calculated the distribution of ΔG values for random protein sequences and found that below the value for $\Delta G_{crit} = 0.0$, only .01% of random sequences are stable. This is concordant with recent experiments demonstrating that most random biological sequences are unfoldable. As such, we expect that sequences are poorly connected in point mutation space. It has been shown that the presence of low connectivity in such protein lattice model spaces establishes neutral nets confined to a single or small set of structures[11]. The primary effect of these walks is to bias sequence space sampling so as to produce heterogeneous substitution rates, which also imply a more biologically relevant model.

3.2 Structures with different designabilities have different rates of structural decay.

Figure 1 emphasizes two quantities important to our understanding of gene duplication. Figure 1a shows the average contact similarity per generation between the foldable protein structures present in the population to the original structure. Figure 1b shows the average designability of the foldable members of a generation. Both of these figures were averaged over 8 different seed sequences in each

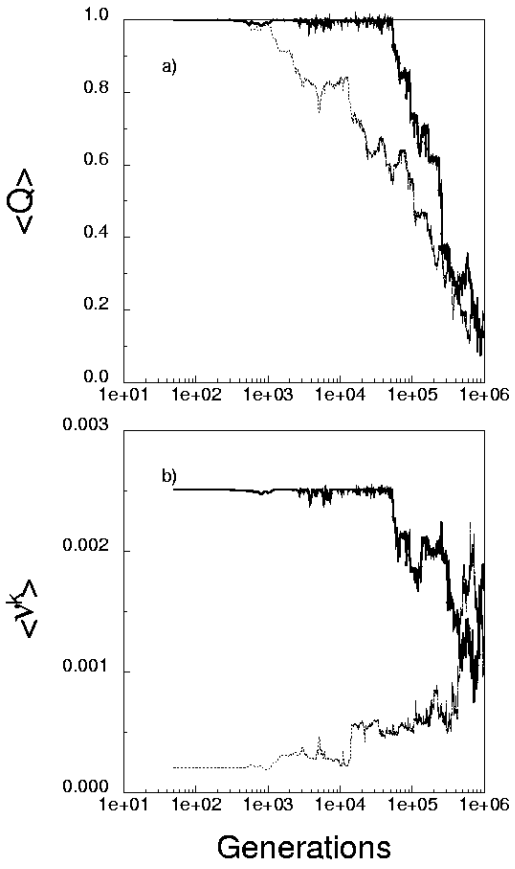

Figure 1: (a) Distribution of average structural similarity <*Q*> for high designable initial structure (-----) and low designable initial structure (…..). (b) Distrubution of average designabilities ν^k for high designable initial structure (-----) and low designable initial structure (…..).

structure constrained by $\Delta G_{crit} <$ 0.0. We now use our results in conjunction with Kimura's neutral theory arguments to discuss the paper's focus.

4. Discussion

In nature, many duplicated genes are able to find functions before the accumulation of mutations in the expressed protein sequence render it unfoldable. In our model, the probability of a single gene mutating to produce an unfoldable sequence (P(*U*)) when ΔG_{crit} = 0.0 is very sensitive to a protein's primary sequence, but usually close to 80%.

4.1 Modest selection pressure for protein stability buffers structural information content of the duplicated gene.

Kimura states that the expected time for a population to be replaced by a neutral mutation is approximately equal to the inverse of the neutral mutation rate[9]. If no selection pressure existed for folding, we might expect our entire population to become unfoldable in about 10^2 generations. A tournament selection pressure of ω = 1.5 is enough to keep almost all of the proteins in the population folded. This would imply that the dominating process involved in losing the genomic information of duplicated genes is the mutational inactivation of the regulation factors, in effect, producing a pseudogene.

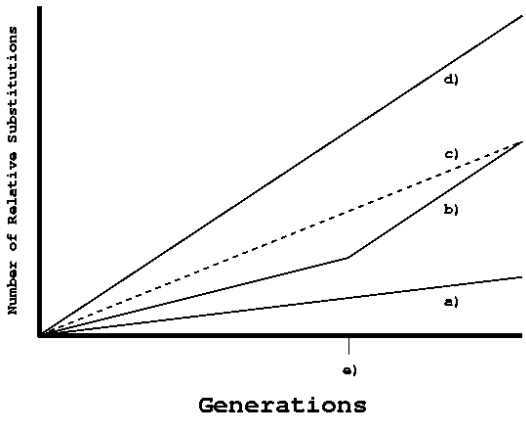

Figure 2: Plot of the number of relative accumulated non-synonymous substitutions over time. (a) Line representing the normal rate of accumulating non-synonymous substitutions *S* if gene is transcribed, translated, and functional. (b) Line representing rates of *S* in the case of a gene duplication whose translated gene product is not functional but must be stable, and becomes a pseudogene at (e). (c) Line representing a possible extrapolated rate *S* from the present day substitution data. (d) Line representing the rate of *S* for a gene with no selection pressure.

4.2 Structural information of highly designable initial structures is preserved in functional duplicated gene products.

In Figure 1a we see that in the case of a low designable structure, the structural similarity is lost after just a few hundred generations. Conversely, the high designable structure conserved its original structure until around 10^5 generations. The significance of this result is amplified by the window of opportunity allowed by the mutational clock affecting the regulation section of the gene. There is an effective limit on the time allowed to search for a new mutation with no selection pressure because of the mutation that might occur in the regulatory region. The function of the protein must be selected for (the evolution must be directed) before said regulatory region is damaged.

Let us assume that the rate of gene deactivation by mutation of the regulation site is a constant random process. Since we are not imposing any selection pressures for functionality, both a foldable mutant and pseudogene have the same selection pressure. Therefore, the expected time for replacement of the population with a pseudogene mutation is approximately equal to the inverse of the regulation-site's deleterious mutation rate. A useful quantity then becomes the ratio of the number of mutations in the protein coding region on the gene to the number of

mutations to the regulation site. If we assume all sites in the gene are equally likely to undergo mutation, then this ratio can be approximated by the relative sizes of protein coding region to regulation region. It seems likely then that the mutation rate of the regulatory region is 10^{-2} times smaller than the protein coding region mutation rate. In our example, the expected number of generations before the population is replaced by pseudogenes is then 10^4 generations. The conservation of the high designable structure is buffered for 10 times longer than the expected production time of the gene. If a selectively advantageous function is discovered in under 10^4 generations, we would most likely expect the product of the high designable gene product duplication to maintain its original structure. The low designable gene product duplication would possibly adopt a novel structure with a slightly higher new designability (as implicated by Figure 1b).

4.3 Selection for weak stability in duplicated gene products might be one reason for the observed relatively rapid evolution in synonymous DNA sites verses the slow evolution of non-synonymous sites in pseudogenes.

In practice it is assumed that on average, the silent sites in pseudogenes reflect the actual substution rate. Both neutralists and selectionists agree that such sites are among the least constrained by evolution in the genome. This provides a method for accurately dating gene family formation events, assuming there exists at least one pseudogene member. Interestingly, the rate of non-synonymous site substution does not seem to behave the same way. Calculated ratios of the rate of the substitution in non-synonymous DNA sites verses that of synonymous sites are often lower than one would expect at random[9,14]. Figure 2 depicts one possible interpretation for the discrepancy. Assuming the molecular clock hypothesis, figure 2 demonstrates the rate of accumulated mutations. There are actually two regimes toward pseudogene formation. An initial slow, structurally conserved process governed by stability of the gene's translated protein product is followed by a second, fast, unconstrained process. The relative ratio is extrapolated back (figure 2, line c) to give a lower estimate of substitution rates in non-synonymous pseudogene sites.

5. Conclusion

We explore issues of gene duplication using a basic model of protein evolution and test the outcomes resulting from the assumption that newly duplicated genes are selected to produce either stable products or become pseudogenes. We find that this should produce a number of observable effects. Firstly, a slight selection pressure for foldable proteins is all that is needed to preserve a population with a majority of stable gene products. This could be tested for experimentally by monitoring the

stability of proteins expressed as the byproduct of an artificial gene duplication. There is some experimental evidence for the existence of natively unfolded protein in organisms with little or no known selection against reproduction efficiency[5,19]. We could account for this in our model since the selection against unfolded states can be tuned by varying the parameter ω. A possible effect of lowered selection pressure for stability might be to allow the genes to search amino acid sequence space more freely for novel functions. However, this would be increasing the risk of having proteins so unstable that it would be difficult for the organism to find a function for them.

Secondly, we predict that structural information is largely conserved in the gene copy before either the discovery of a novel protein function or formation of a pseudogene can occur. A recent theoretical study[13] showed that highly designable structures are more likely to be observed as a direct result of population evolution. As these structures are more probable, and since highly designable structures have a greater chance of preserving structural similarity, we might expect to see an over representation of multi-gene families with similar structures. This could also be tested directly as before, by computing the structural similarity along with the stability of proteins produced by a population of organisms. We suspect that sequences mutating from lower designable structure stand a better chance at developing novel structures based on their connectivity to other structures. Interestingly, however, these will still share some structural similarity to the original structure because the neighbors to low designable structures share more common contacts[20]. Since structure is often related to function, we might also expect a high degree of preserved function among the multi-gene families, although the cause of this might be hard if not impossible to decouple from selection imposed by the organism for gene dosage effects.

Finally, the selected ability to remain stable should effect the way substitution rates are calculated. Conserving structure also biases conservation of amino acid sequence. Only after translation is inactivated would true neutral evolution have a chance to occur. This is also quite testable as a hypothesis. Looking at Figure 2, we can see that relaxing the stability constraint ΔG_{crit} increases the slope of the line. The slope of the line reaches a maximum for completely unconstrained neutral evolution. In the future, we intend to recreate figure 2 as a function of ω using simulations similar to those used in this paper. We also wish to use the homologies of multi-gene families containing pseudogenes to fit rates of non-synonymous substitutions to multiple curves. If the effect of stability is strong enough, we might even be able to produce more accurate timing of phylogenetic events. However, we should be careful of loose interpretations as evidence of repair mechanisms, nucleotide frequencies or tRNA concentrations might influence this effect[6,9].

We understand that biology is very complex and some important issues are not discussed here. For instance, a pair of genes resulting from duplication must

coevolve together under evolutionary pressures. We do not consider the rich complexities allowed by the possibility of feedback regulation, where a gene product can affect it's own transcription (as in the case of the lac operon). Also, the existence of neutral nets is a key factor in the sampling of the sequence space. To what extent do such nets and selection pressures exist in real proteins? In addition, there are repair mechanisms in the cell that use one gene as a template to repair another. This gene conversion may very well be the dominant reason for pseudogene sites conservation ratios[2]. Also, these same specialized repair mechanisms may be responsible for prolonging the lifespan of the regulatory region of the gene, effectively allowing it to produce longer than would usually be allowed in the presence of normal mutation.

Acknowledgments

We would like to thank Matt Dimmic and our anonymous reviewers for helpful comments. Financial support was provided by NIH grants GM08297 and LM0577.

References

1. T. Ohta, *PNAS* **85** 3509 (1988).
2. W.-H. Li *et al*, *Fundamentals of Molecular Evolution*. (Sinauer Associates, Inc., Massachusetts, 1991).
3. A. Wagner, *J. Evol. Biol.* **12** 1 (1999).
4. J.B.S. Haldane, *Am Nat.* **708** 5 (1933).
5. F.J. Blanco *et al*, *JMB* **285** 741 (1999).
6. Nathalie Trabesinger-Ruef *et al*, *FEBS* **382** 319 (1996).
7. A. Wagner, *BioEssays* **20** 785 (1998).
8. M.A. Nowak *et al*, *Nature* **388** 167 (1997).
9. M. Ridley, *Evolution 2nd ed.* (Blackwell Science, Cambridge, 1996).
10. B.J. Walsh, *Genetics* **139** 421 (1995).
11. S. Govindarajan *et al*, *Proteins* **29** 461 (1997).
12. N.E.G. Buchler *et al*, *J. Chem. Phys.* **In press**.
13. D.M. Taverna *et al*, *Biopolymers* **In press**.
14. A. Force *et al*, *Genetics* **151** 1531 (1999).
15. S. Miyazawa *et al*, *Macromol*. **18** 534 (1985)
16. S. Govindarajan *et al*, *PNAS* **95** 5545 (1998).
17. M. Kimura *et al*, *PNAS* **76** 2858 (1979).
18. T. Ohta, *Genetics* **115** 207 (1987).
19. P.H. Weintreb *et al*, *Biochemistry* **35** 13709 (1996).
20. S. Govindarajan *et al*, *PNAS* **93** 3341 (1996).