Pacific Symposium on Biocomputing 4:90-101 (1999)

EVOLUTIONARY CONSTRAINT NETWORKS IN LIGAND-BINDING DOMAINS: AN INFORMATION-THEORETIC APPROACH

SYLVIA B. NAGL*, JAMES FREEMAN and TEMPLE F. SMITH

Biomolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street, Boston, MA 02215.
E-mail: nagls@darwin.bu.edu
jfreeman@darwin.bu.edu
tsmith@darwin.bu.edu

Ligand-binding sites in homologous protein domains can diverge greatly during evolution. This poses a particularly interesting problem in those cases where the ligand-binding site is situated in, or close to, the domain core, or where ligand-docking induces dramatic conformational changes. These features are present in many receptors and enzymes; the hormone-binding domain of the nuclear receptors for steroids and retinoids, for example, exhibits both characteristics. It is therefore of great interest to determine how binding sites for diverse ligands evolve in core regions of structurally dynamic domains. Are evolutionary changes locally restricted to the ligand-binding site, or are they distributed throughout the domain? We describe here an information-theoretic approach for the study of covariation between ligand-contacting residues and compensatory mutations that preserve the structural integrity and the conformational dynamics of ligand-binding domains. We apply this method to the analysis of the nuclear receptor ligand-binding domain and show that the ligand-contacting residues in the hormone-binding pocket are evolutionarily linked to an extensive network of covarying positions.

1 Introduction

Several features make the ligand-binding domain in nuclear receptors a prime example for the evolution of binding sites in core regions of structurally dynamic domains. Firstly, steroid, thyroid and retinoid hormones comprise the most diverse class of gene-regulatory ligands known (1, 2). Their receptors belong to the superfamily of nuclear receptors (NRs) present in all metazoans (3). As ligand-inducible transcription factors, NRs play essential roles in the regulatory pathways that transmit signals, originating from the extra- and intra-cellular environment, to large genetic networks through a complex sequence of molecular interactions. Secondly, crystallographic studies of five NR ligand-binding domains (LBDs) suggest a structural role for the ligand that is fundamental to LBD function (4-8). Approximately 65% of the LBD domain is α helical; all five LBDs share a prototypic fold that has been termed an “antiparallel α helical sandwich”. The helices are grouped into three layers around an internal ligand-binding core. The ligand is completely buried within the domain interior and contributes to the hydrophobic core of the active conformation of the NR. Ligand binding directs the alignment of the secondary structural elements and strongly constrains the conformational freedom of the LBD. Contacts with the ligand are extensive and include at least eight different structural elements throughout the length of the LBD. Thirdly, within the structural constraints of the LBD core, ligand-binding pockets for ligands possessing strikingly diverse chemical structures evolved independently in different NR families (1, 3). Escriva *et al.* (1997) proposed that the ancestor of the superfamily was an orphan receptor, possibly an RXR-like protein, without ligand-binding capability (3). Since the ligand plays a structural role, several potentially conflicting constraints have to be satisfied during LBD evolution: (i) spatial accommodation of the ligand within the core; (ii) high affinity and specificity of ligand binding; (iii) maintenance of overall LBD structural integrity; (iv) conservation of LBD dynamics (allosteric controls); and (v) formation of functional surfaces.

Our information-theoretic study identified significant covariation between ligand-contacting residues and correlated positions located in various regions of the LBD. The nature of the mutations in correlated positions suggests that they compensate for the binding of diverse ligands and preserve the structural integrity and the conformational dynamics of the LBD. We anticipate that our novel approach to the analysis of ligand-binding sites in protein (super)families will find wide-ranging applications in molecular modeling and design. This new approach is universally applicable to protein targets of pharmaceutical interest. These include not only receptors for steroid/retinoid hormones, but also receptors for neurotransmitters, signal molecules of the immune system, and growth factors.

2 Materials and Methods

2.1 Data set and alignment procedures

Sequences from 51 members of the nuclear receptor superfamily, representative of the entire superfamily, were retrieved from Translated Genbank. Only a single sequence was included from each subfamily of closely related sequences. A structurally constrained, pattern-induced multi-sequence alignment (PIMA) (9) of the sequence regions homologous to the prototypic fold of the LBD (4-8) was performed through iterative pairwise local dynamic programming. Sequences were aligned from helix 3 onward, since the N-terminal portion of the LBD is disordered and a reliable alignment can not be obtained due to low sequence similarity. The crystallographic structure of the human RARγ LBD (PDB code 2lbd.ent) (6) was used as a template to create a structurally constrained sequence alignment that allowed gap placement to be almost exclusively restricted to known loop regions (10). Alignment positions were numbered according to their homologous residues in the human RARγ LBD. The LBD alignment file is available upon request.

2.2 Analysis methods

The concept of a ligand-binding space, as an abstract representation of all possible ligand interactions the LBD can engage in within the structural constraints of its prototypical fold, is useful for the investigation of the evolution of ligand-binding in NRs. Within this conceptual framework, the emergence of new ligand-binding capabilities can be understood as the result of adaptive walks in LBD sequence space. During this adaptive evolution, NR genes accumulate successive mutations that progressively increase the affinity of the ligand-binding pocket for new ligands.

From an evolutionary perspective, amino acid positions in the LBD can be classified in four groups. These are: positions with conserved amino acid identities, positions with conserved physicochemical properties, positions with highly variable physicochemical properties, and unconstrained positions accumulating neutral mutations. With the exception of neutral positions, each amino acid residue makes an individual fitness contribution and simultaneously affects the fitness of *n* other residues within the LBD. Fitness is here defined as the capacity of the LBD to maintain its structural and functional features and to bind a specific ligand.

We identified correlated positions from the aligned LBD sequences using mutual information, a measure of correlation for discrete symbols (11). Columns in the alignment that were primarily gapped to maintain alignment with sequences that contain amino acid insertions were deleted. A formal measure of variability at position i is the Shannon entropy, $H(i)$. $H(i)$ is defined in terms of the probabilities $P(s_i)$, of the different symbols, s, that can appear at a sequence position (i.e., for amino acid sequences s = 20, for the 20 possible states of amino acid occurrence) (12). In this study, the observed amino acid frequencies were used, since the true probabilities of the population are unknown. $H(i)$ is defined as

$$H(i) = -\sum_{s} P(s_i) \log P(s_i) \quad (1)$$

Mutual information is defined in terms of entropies involving the joint probability distribution, $P(s_i,s'_j)$, of occurrence of symbol s at position i, and s' at position j. The associated entropies for each position i and j are

$$H(i) = -\sum_{s_i} P(s_i) \log P(s_i) \quad (2)$$

$$H(j) = -\sum_{s'_j} P(s'_j) \log P(s'_j) \quad (3).$$

And the joint entropy is defined as

$$H(i, j) = -\sum_{s_i, s'_j} P(s_i, s'_j) \log P(s_i ,s'j) \quad (4).$$

The mutual information, $M(i, j)$, is defined as

$$M(\mathrm{i}, \mathrm{j}) = H(i) + H(j) - H(i, j) \quad (5).$$

If the positions are independent, their mutual information is 0. If, on the other hand, the positions are correlated, their mutual information is positive and achieves its maximum value if there is complete covariation.

2.3 Determination of the significance of the mutual information scores

Given a set of sequences that are assumed to be independent and identically distributed samples from a probability distribution, one can independently estimate each pairwise probability distribution for every pair of positions by frequency counting. However, sequences belonging to a protein (super)family are not independent samples, but are related through shared ancestry described by a phylogenetic tree. To estimate the mutual information content between position pairs that is created by tree inheritance alone we performed a control simulation experiment. Our procedure was based on the null-model method described by Lapedes *et al*. (1995) (13) with modifications. The experiment simulated the evolution of sequences by random mutations along a phylogenetic tree obtained from the 51 nuclear receptor LBD sequences by maximum parsimony. A PAUP tree was constructed with a heuristic search algorithm and 100 bootstrap replicates using the GCG Seqlab package (tree available on request). All characters were weighted equally, and all were parsimony-informative. Using the outgroup (LBD sequence of *C. elegans* NR CeF11C1) as a seed, random sequences were evolved following the PAUP phylogenetic tree obtained from the real data set. During simulated random mutation of sequences, the state of the sequence was duplicated at a bifurcation point in the tree, and the two copies were then independently evolved. Every amino acid could mutate with equal probability to any other amino acid. Indels were not considered in the mutations. The procedure was repeated numerous times, and significance threshold values were thereby determined from the frequency distributions of the mutual information scores in the control and real data sets.

3 Results and discussion

3.1 Identification of LBD residue positions that covary with ligand-contacting positions

Hormone recognition is achieved through a combination of specific hydrogen bonds and the complementarity of the binding cavity to the non-polar ligand. Ligand-contacting residues (within 4.5 C of the ligand) have been identified by X-ray crystallography for the retinoic acid receptor (RAR) (6), the thyroid hormone receptor (TR) (5), the estrogen receptor (ER) (7), and progesterone receptor (PR) (8). Ligand contacts have also been modeled for the vitamin D receptor (VDR) (14). The ligand-binding characteristics of this small set of receptors are of necessity a biased sample of all ligand-interactions NRs are engaged in. Nevertheless, an alignment of the sequences of this set allows the identification of 28 homologous residue positions that are involved in binding several structurally diverse ligands, as well as 12 ligand-specific contact positions (data available at http://bmerc-www.bu.edu/nagls/PSB99). We determined the mutual information between those positions in the LBD sequence alignment that are homologous to ligand-contacting residues in one or more of the receptors for which crystallographic data are available and all other positions in the LBD. The complete set of significant correlations can be obtained at http://bmerc-www.bu.edu/nagls/PSB99.

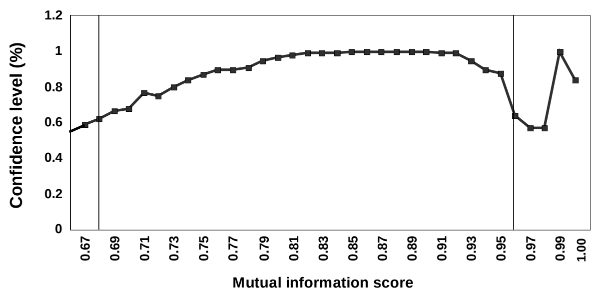

Figure 1. Confidence levels for mutual information scores. In order to control for small sample bias and phylogenetic effects, simulated evolution of random sequences along a maximum parsimony tree for the LBD was performed repeatedly. Confidence limits for the mutual information scores were determined from the frequency distributions of control and real data sets. Scores between 0.68 and 0.96 have a greater than 60% probability (*confidence level*) of not being caused by phylogenetic effects (see threshold lines). Scores > 0.98 are caused by invariant positions in the data set and were therefore excluded (see text for details).

In order to control for small sample bias and phylogenetic effects, simulated evolution of random sequences along a maximum parsimony tree for the LBD was performed repeatedly, as detailed in Materials and Methods. Confidence limits for the mutual information scores were determined from the frequency distributions of the mutual information scores in the control and real data sets (Fig. 1). It was shown that scores between 0.68 and 0.96 (inclusive) have a 60-100% probability of not being caused by phylogenetic effects in a small sample population. For the purpose of the present study, we chose a confidence limit of 60% to make certain that the full range of potentially significant correlations is detected. The absence of any prior knowledge about the domain constraints within which the evolution of ligand-binding sites has occurred necessitates this approach. The upper threshold excludes highly conserved positions that create high mutual information scores with *any* other position in the absence of covariation (>50% amino acid identity).

3.2 Ligand pocket residues are correlated with a network of covarying positions located throughout the LBD

Since the ligand-contacting residues line the hormone-binding pocket in the domain core, they perform a dual role; a functional role in ligand recognition and a structural role as core residues. With respect to ligand recognition, they can be seen to constitute an 'interior interaction surface'. In agreement with this interpretation, we found that contact positions within the ligand-binding pocket do not constrain each other's amino acid variability (no significant correlations). In principle, this would allow extraordinary scope for the evolution of the ligand-binding pocket. However, since the hydrophobic ligand is an integral part of the LBD core in the active conformation, the ligand and the ligand-binding residues combined need to be able to maintain structural stability and domain dynamics (conformational changes). How is this potential conflict between structural constraints and functional diversity resolved within the LBD fold?

This conflict primarily concerns volume and shape changes in the domain core due to structural differences between ligands (some of these may be offset by a conformational change in the ligand upon binding). In addition to exhibiting great structural diversity, the van der Waals volumes of the eleven known and seven putative NR ligands are between 201.3 and 426.8 C^3 (15). It remains to be determined whether the volumes of other, as yet unknown, natural NR ligands also fall within this already considerable range, or show even greater differences. In general, structural changes, including those caused by the binding of diverse ligands in core regions, are limited by the low free energy of proteins and the need to maintain thermodynamic stability. If an energetically unfavorable change occurs, complementary mutations are needed to restore the previous structure or to stabilize the structural change. These changes might be compensated for in two ways. Firstly, mutations in individual contact residues might directly alter the shape of the binding surface. Secondly, mutations at helix interfaces throughout the LBD that change the relative positions and orientations of the helices might indirectly accommodate volume changes in the ligand pocket.

The positions found to covary significantly with ligand-contacting residues, and implicitly, with different ligands, map to helical regions in the LBD, namely helices 3-4, 7-10 and 12 in the hRAR γ LBD crystal structure (data available at http://bmerc-www.bu.edu/nagls/PSB99). An overall covariation pattern is apparent which is characterized by multiple correlations between two *covarying sets* of positions, the ligand contacting residues and a distributed covarying network (Fig. 2). This distinctive pattern reflects the structure-function relationships within the LBD. The first group, made up of the ligand-contacting positions, functions as an integrated set in the recognition of a specific ligand. The second set, the covarying network as a whole, is proposed to compensate for volume changes in the domain core that are caused by the binding of diverse ligands. Because of the flexibility of protein structure, the compensatory response may be distributed over a cluster of residues. The fact that LBDs of nuclear receptors are highly divergent supports this hypothesis. Considering volume compensation, with greater evolutionary distance, a greater number of substitutions occur, increasing the probability that the volume compensation necessary to preserve fold integrity will be achieved by substitution at multiple positions. This compensatory process would be expected to occur in parallel with the accumulation of successive mutations in the ligand-binding pocket. Such an integrated adaptive walk in LBD sequence space would progressively increase the affinity of the ligand-binding pocket for new ligands.

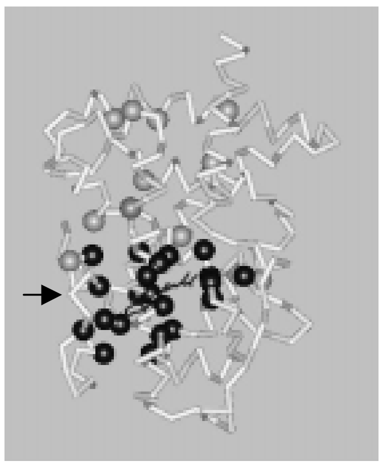

Figure 2. Ligand-contacting positions and the covarying network of the RAR LBD. Ligand contacts are shown in *black*, covarying positions are shown in *grey* (α–carbons, spacefill mode). The ligand is shown in *black* (stick mode). Helix 12 is marked bv an arrow.

Changes in core volume are highly amenable to this mode of compensatory evolution (consider, in contrast, a loss of charge which could not be compensated for by a gain of two half charges at two other positions). Core volume changes are accompanied by changes in the geometry of the helix packings (16). Proteins accommodate mutations in core regions by a change of structure which principally involves rigid-body movements of helices relative to each other (up to 7 and 30^{o}) and dissipation of the movement in the turn regions. Since specific interactions among side chains dictate the relative orientations of secondary elements, a distributed network of covarying positions could provide a fine-tuning mechanism for maintaining structural integrity.

As shown in Fig. 3A, the covarying network is closely linked to the highly conserved signature motif region (17), which spans helices 3 and 4 and stabilizes the LBD fold. Three covarying positions are located adjacent to conserved motif residues (RAR242-$\underline{P}_c$xMMMxMM$\underline{P}_c$MxxMxxxMMxxMM$\underline{P}_c$-264RAR; $\underline{P}_c$, covarying position [*black* in Fig. 3A]; M, motif position [*grey* in Fig. 3A]; x, intervening position). This suggests that, interestingly, covarying positions provide a mechanism for structural flexibility in juxtaposition to constrained elements that conserve the pattern of the fold. These covarying positions are, in turn, linked to a third 'tier' of covarying positions. Fig. 3B illustrates the hierarchical structure of the part of the covarying network that is centered around the signature motif region. Statistically significant values of mutual information typically identify complicated hierarchical 'chains of correlations' between sites that are not physically in contact (12, 13). These may point to the presence of higher-order interactions between sites; for instance, in the case of two unconnected sites that are both in contact with a third site ('shared causation versus covariation'; 13). It is still an open question whether the networks of correlations identified in the LBD of nuclear receptors constitute such cases, or whether they covary at a distance. The hypothesis of an evolutionary mechanism for structural integrity in ligand-binding domains that involves distributed sites at which selection occurs in parallel would make covariation at a distance conceivable.

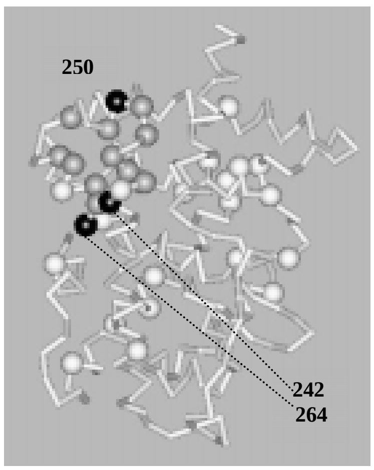

Figure 3A. The covarying network is linked to the signature motif region. Positions that covary with ligand contacts are shown in *black* (*numbered*), and their correlated positions are shown in *white* respectively. Signature motif positions are shown in *grey*.

Positions in the covarying network also play an essential role in LBD conformational dynamics. Significant covariation is present between contact residues in the ligand-binding pocket and six residues in the AF-2 AD core within the transactivation helix 12 (RAR 408, 410, 411, 414-416) (Fig. 2; http://bmerc-www.bu.edu/nagls/PSB99). Positions 408, 411, and 415 are themselves ligand contacts. In the active conformation of the RAR LBD, helix 12 is tightly packed against the body of the domain, and the solvent-exposed residues of the AF-2 AD core are involved in transactivator binding. The residues on the opposite side of the amphipathic helix stabilize the active conformation. In the activated conformation, covarying positions RAR 408, 415 and 416 are oriented toward the domain body and

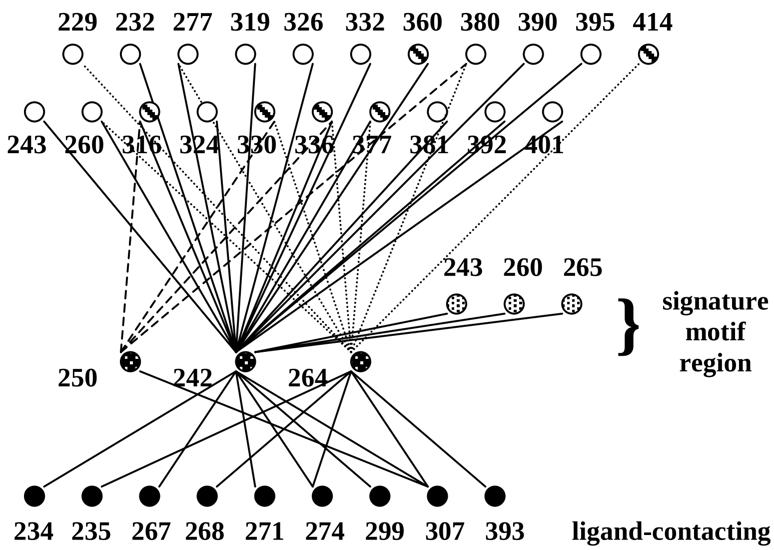

Figure 3B. The hierarchical structure of the part of the covarying network that is centered around the signature motif region. Ligand contact positions are shown as ●, positions in the covarying network located in the signature motif region are shown as ❂, and positions within a second 'tier' of covariation are shown as ◌ (in the motif region), as ⊘ (also in covarying network), or as ○.

contribute to the positioning of helix 12. In some NRs for which crystal structures are not available, residue locations in helix 12 might be shifted relative to those in RAR, and residues 410 and 411 might interact with the domain body.

RAR E414 forms a salt bridge with K264 in helix 4 in an electrostatic interaction which anchors helix 12 (6). Residues 414 and 264 covary with each other in NRs. Charge reversals preserve the electrostatic interaction in several receptors, but the salt bridge is not conserved in all NRs. Surprisingly, mutually repulsive charges (notably E-E) are present in numerous NRs. These would be expected to destabilize the canonical active conformation, so far observed in all LBD crystal structures. This change might either contribute to a structural modification to accommodate a certain type of ligand and the creation of a new interaction surface for transactivators, or might result in a loss of transactivation and a dominant negative phenotype.

3.3 Several clinically important mutations in nuclear receptors are located in the correlated network

Finally, our study shows that several dominant negative mutations in the human androgen receptor (AR), estrogen receptor (ER), and thyroid hormone receptor β (TRβ), previously shown to result in generalized hormone resistance or loss of transactivation, are located in the network of covarying positions (Table 1). This suggests that, in some cases, the information-theoretic approach described in this paper can contribute to the study of single-site polymorphisms implicated in differential responses to hormones or drugs in humans.

4 Conclusion

We showed that ligand-contacting positions in the hormone-binding pocket of nuclear receptors are evolutionarily linked to a correlated network of positions located throughout the LBD. Ligand-responsive nuclear receptors appear to have evolved from ancient orphan receptors that could assume an active conformation in the absence of ligand-binding (18). This conformational versatility is an extraordinary feature of the LBD, and the ligand is thought to alter an equilibrium between multiple inactive and active states (19-21). Recent work on antibody maturation suggests that alternative conformations of proteins may have an evolutionary role similar to gene duplications (22). Wedemayer *et al.* suggested a process whereby one conformation maintains structural and functional fitness, and alternative conformations may evolve new functions. Moreover, they showed that the mutational events driving this process occur more frequently at positions that are distant from the antigen-binding site. This mutational pattern is similar to the evolutionary process involving the covarying networks in the LBD. In analogy to germline antibodies which evolve to bind an almost limitless array of potential antigens, evolution of diverse ligand-binding capabilities in the LBD of NRs may have been made possible by the striking conformational versatility of this domain. The conformational flexibility of the LBD might in fact contribute to its evolvability.

Table 1. Dominant negative mutations in the correlated network. A large number of mutations in the LBD of steroid and thyroid hormone receptors have been previously identified that result in a dominant negative phenotype. Some of these mutations lead to a loss of hormone binding, or loss of transactivation, although they do not form part of the ligand pocket. Listed here are dominant negative mutations that are located at positions that covary with ligand-contacting residues.

NR	mutation	RAR_{equiv}[a]	ref.	covarying ligand contact(s)	RAR_{equiv}[a]
ER	C447A	330	544257[b]	ER 347	231
				ER 354	238
AR[c]	M807R	331	113830[b]	AR 705	231
				AR 707	233
				AR 780	305
				AR 877	396
				AR 894	411
				AR 895	412
				AR 898	415
AR[c]	D864N	383	113830[b]	AR705	231
				AR707	233
				AR784	308
				AR787	311
TRβ1	P453T	410	586092[b]	TRβ1454	411

[a] Homologous position in RAR (crystallographic structure available, PDB code 2lbd.ent).
[b] GENBANK gi number, see annotation.
[c] Ligand contact residues not confirmed by X-ray crystallography. Covarying positions in the ligand pocket correspond to ligand contacts in ER, with the exception of RAR_{equiv} 415.

Acknowledgments

We thank the staff at the BMERC for helpful discussions. This research was supported in part by a postdoctoral fellowship grant from the Foundation of Research, Science and Technology, New Zealand (to S. B. N.).

References

1. D. J. Mangelsdorf, et al., *Cell* 83, 835-839 (1995).
2. V. Laudet, et al., *EMBO Journal* 11, 1003-1013 (1992).
3. H. Escriva, et al., *Proc. Natl. Acad. Sci.USA* 94, 6803-6808 (1997).
4. W. Bourguet, et al., *Nature* 375, 377-383 (1995).
5. R. L. Wagner, et al., *Nature* 378, 690-697 (1995).
6. J. P. Renaud, et al., *Nature* 378, 681-689 (1995).
7. A. M. Brzozowski, et al., *Nature* 389, 753-758 (1997).
8. S. P. Williams and P. B. Sigler, *Nature* 393, 392-396 (1998).
9. R. F. Smith and T. F. Smith, *Proc. Natl. Acad. Sci. USA* 87, 118-122 (1990).
10. R. F. Smith and T. F. Smith, *Protein Eng.* 5, 35-41 (1992).
11. T. Cover and J. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications (John Wiley and Sons, New York, 1991).
12. B. T. M. Korber, et al., *Proc. Natl. Acad. Sci. USA* 90, 7176-7180 (1993).
13. A. S. Lapedes, et al., AMS/SIAM Conference on Statistics in Molecular Biology, Seattle, WA (1997).
14. J.-M. Wurtz, et al., *in* A. W. Norman, R. Bouillon, M. Thomasset, Eds., Vitamin D: Chemistry, Biology and Clinical Applications of the Steroid Hormone, Strasbourg, France (U. Cal. Riverside, 1997).
15. A. A. Bogan et al., *Nature Structural Biology* 5, 679-681 (1998).
16. A. M. Lesk and C. Chothia, *J. Mol. Biol.* 136, 225-270 (1980).
17. J.-M. Wurtz, et al., *Nature Structural Biology* 3, 87-94 (1996).
18. H. Escriva, et al., *Proc. Natl. Acad. Sci. USA* 94, 6803-6808 (1997).
19. S. B. Nagl, et al., *Mol. Endocrinol.* 9, 1522-1532 (1995).
20. S. B. Nagl, et al., *J. Cell. Biochem.* 67, 184-200 (1997).
21. J. A. Katzenellenbogen, et al., *Mol. Endocrinol.* 10, 119-131 (1996).
22. G. J. Wedemayer, et al., *Science* 276, 1665-1669 (1997).