Sequence annotation of nuclear receptor ligand-binding domains by automated homology modeling

C.Jan J.François, Jan P.G.Klomp and Ronald M.A.Knegtel

1Laboratory of Biochemistry, Wageningen Agricultural University, Dreijenlaan 3, 6703 HA Wageningen, 2Target Discovery Unit and 3Department of Molecular Design and Informatics, NV Organon, P.O. Box 20, 5340 BH Oss, The Netherlands

To whom correspondence should be addressed. Present address: Vertex Pharmaceuticals (Europe) Limited, 88 Milton Park, Abingdon, OX14 4RY, UK. E-mail: rknegtel@vpharm.com

The quality of three-dimensional homology models derived from protein sequences provides an independent measure of the suitability of a protein sequence for a certain fold. We have used automated homology modeling and model assessment tools to identify putative nuclear hormone receptor ligand-binding domains in the genome of Caenorhabditis elegans. Our results indicate that the availability of multiple crystal structures is crucial to obtaining useful models in this receptor family. The majority of annotated mammalian nuclear hormone receptors could be assigned to a ligand-binding domain fold by using the best model derived from any of four template structures. This strategy also assigned the ligand-binding domain fold to a number of C.elegans sequences without prior annotation. Interestingly, the retinoic acid receptor crystal structure contributed most to the number of sequences that could be assigned to a ligand-binding domain fold. Several causes for this can be suggested, including the high quality of this protein structure in terms of our assessment tools, similarity between the biological function or ligand of this receptor and the modeled genes and gene duplication in C.elegans. Keywords: automated homology modeling/Caenorhabditis elegans/fold recognition/nuclear receptors/sequence annotation

Introduction

The nuclear hormone receptor family constitutes an important target for drug discovery (Beato et al., 1995; Kliwer et al., 1999). These receptor proteins bind directly to specific DNA recognition sequences in the regulatory regions of genes, resulting in the alteration of gene transcription and translation. The family includes receptors for steroid and thyroid hormones, vitamins, retinoids and prostaglandins. Synthetic ligands for this protein family currently find applications in therapeutic areas such as human fertility, osteoporosis, oncology, dermatology and cardiovascular diseases. For only a limited set of receptors, however, have the biological role and native ligands been determined. The remaining proteins are commonly referred to as ‘orphan receptors’, reflecting the unknown nature of the endogenous ligand (Kliwer et al., 1999). Given the successful commercialization of synthetic ligands for receptors of known function, the discovery of novel hormone receptors is of high interest to the pharmaceutical industry. In addition, the identification of nuclear receptors in lower species can be used to elucidate regulatory pathways that may have homologues in humans.

Nuclear hormone receptor genes are commonly identified and annotated on the basis of the highly conserved, N-terminal DNA binding domain (DBD) (Clarke and Berg, 1998). Not every hormone receptor contains a DBD, however, and not every protein containing a C4-type Zinc-finger DBD also features a hormone-binding domain. In addition, reliable identification of the presence of a ligand-binding domain (LBD) is complicated by the relatively low sequence identities observed among LBDs. For instance, annotations for 233 nuclear receptors in Caenorhabditis elegans were obtained using hidden Markov models (HMMs) for the DBD (Clarke and Berg, 1998). Only after applying hidden Markov models for the LBD, eventually incorporating also C.elegans sequences that scored well in previous searches, could an LBD fold tentatively be assigned to 90% of the predicted proteins. The remaining 10% of the predicted proteins with a DNA-binding domain remained without further annotation for a ligand-binding domain.

When sequence identity alone does not allow for the unambiguous annotation of a gene, a numerical estimate of the compatibility of a sequence with a three-dimensional (3D) protein fold provides an independent measure for annotating sequences. Several different approaches to obtaining such an estimate have been described and are used in fold recognition software (Torda, 1997). If a 3D model can be obtained for a protein sequence, model validation tools can be used to assess its suitability for a particular fold.

Homology or comparative modeling allows for the automatic generation of 3D structural models for protein sequences by using the structure(s) of sequence homologues as templates. Homology modeling can be expected to yield meaningful 3D models for sequence–template pairs with sequence identities over 30%. Sánchez and Sali (1998) followed such a strategy for gene annotation by generating homology models for all sequences in the Saccharomyces cerevisiae genome that shared sequence homology with structures in the protein databank (PDB).

As the basic fold of the nuclear hormone receptor LBD has been known since the elucidation of several structures within this family, homology modeling is an attractive technique to generate and evaluate models for sequences of putative receptors in C.elegans. Although the sequence identity among hormone-binding domains of nuclear receptors is often below 20%, the strong conservation of the adjacent N-terminal DNA binding domain allows for the localization of the putative hormone-binding domain and its alignment with other sequences.

Here we present results on automated homology modeling and model assessment for a database consisting of known and (partially) annotated nuclear receptor sequences from mammals and C.elegans, respectively. The mammalian sequences serve both as an internal control and to calibrate the criteria for annotation of non-mammalian sequences. We discuss the importance of using different LBD crystal structures for model comparison.
building and their role in the annotation of putative hormone receptor sequences.

Materials and methods
To date, the nuclear hormone receptor family consists of at least 50 different members in humans and about 85 different members if lower eukaryotes are also taken into account (all sequences were retrieved from Swissprot, PIR, Genbank and Trembl). The genomic sequence of the nematode Caenorhabditis elegans reveals more than 19,000 predicted genes, of which approximately 270 contain DNA binding C4 Zinc-finger domains. A database was created consisting of protein sequences from the previously mentioned selection of 85 different nuclear receptor family members and 233 predicted C.elegans proteins containing Zinc-finger domains, as used by Clarke and Berg (1998).

Multiple sequence alignments were produced with ClustalW and its graphical interface ClustalX using default parameters. First, all sequences in the database were aligned in a multiple alignment. Sequences that caused the insertion of large gaps in the Zinc-finger domain region were removed. Sequences that did not contain Zinc-finger motifs were kept if they did not disrupt the alignments. The portion of aligned sequences C-terminal of the Zinc-finger domains was saved for further alignment.

Subsequently, a profile of annotated and aligned nuclear hormone receptors was created from human, Drosophila melanogaster and C.elegans nuclear receptor LBD sequences. This profile was used to align the LBD sequences in our database. Sequences that disturbed the N-terminal region of the LBD sequences, containing the strongly conserved F/WAK tripeptide sequence, were removed. The resulting alignment was trimmed to delete all amino acids before and after the ligand-binding domains as delimited by the crystal structures. Finally, a profile of the sequences of the four known structures was created and the remaining sequences (206) were individually aligned with this profile.

Nuclear hormone receptor crystal structures were chosen from those available at the initiation of this project in the PDB. For the estrogen receptor (ER), progesterone (PR), retinoic acid-γ (RAR) and retinoid X-α receptor (RXR) ligand-binding domains PDB entries 2ERD (Shiau et al., 1998), 1A28 (Williams and Sigler, 1998), 2LBD (Renaud et al., 1995) and 1LBD (Bourguet et al., 1995) were used, respectively. The pairwise sequence identities of these LBD sequences were in the range 14–31%. Using different structures for the same protein as templates (which are, for instance, available for the ER) did not yield significant differences in Z-scores. Therefore, only one representative structure was used per receptor. All ligands and waters were removed prior to their application as templates for homology modeling. Models were built using Modeler 5.0 (Sali and Blundell, 1993) as distributed with Quanta 98 (MSI, San Diego, CA) using default parameters and no further optimization. The ‘very fast’ and ‘Align2D’ options in Modeler 5.0 were used with default parameters (Sanchez and Sali, 1998). These options proved useful for speeding up model generation by a factor of 2 and utilize template-derived secondary structure information to improve the initial alignments, respectively. The Z-score as calculated by ProsaII version 3.0 (Sippl, 1993) was used to assess the quality of the models. No correction was made for differences in sequence length, as these were small after trimming the sequences to the length of those of the crystal structures. In our hands, the ProsaII Z-score performed better in discriminat-

ing true hormone receptors from unrelated sequences in a comparison with the Lüthy–Eisenberg 3D/1D profile (Lüthy et al., 1992) and the packing quality assessment of WhatCheck (Vriend et al., 1993). All further analysis was performed with in-house developed software, Quanta 98 and Excel (Microsoft).

PSI-BLAST (Altschul et al., 1997) sequence alignment searches were performed for comparison using Blastp 2.0. Four Blast searches were performed using the four template sequences individually against the LBD database to which the template sequences had been added. This allows for the automatic construction of a profile based on the four template sequences by PSI-BLAST, similar to the use of a profile in aligning sequences for model generation. For each template sequence three iterations were performed and profiles were accumulated from all sequences with E-values smaller than 10^{-7}. The final selection of sequences identified by PSI-BLAST to be similar to the template LBD were those with E-values smaller than 10^{-4}. All sequences were subsequently collected and reduced to a list of unique sequences with reasonable similarity to the four templates.

Results and discussion
Prior to the generation of homology models, sequence alignments were further refined with the Align2D option implemented in Modeler 5.0 (Sánchez and Sali, 1998). This option takes the location of secondary structure elements in the template structure into account while placing insertions. Models based on pairwise or profile alignments without further refinement with Align2D had less favorable Z-scores than those obtained with Align2D (results not shown). In addition, the rankings of the Z-scores of receptor models obtained using different templates, as assessed with the Spearman rank correlation coefficient r (Lehmann and D’Abrera, 1998), are more similar when only the profile-based alignment was used (r = 0.81–0.88) than when Align2D was also applied (r = 0.34–0.80). The Spearman rank correlation coefficient provides a measure of the similarity of two ranked lists (of sequences in the present case). The correlation coefficients for the profile-derived rankings are mutually more similar and closer to 1 than those for the rankings obtained through the application of Align2D. Since the Align2D procedure uses secondary structure information from the individual template structures in improving sequence alignments, it is not unexpected that alignments and Z-score rankings of models depend on the template structure that is used to derive them. In contrast, alignments and models derived using only a single sequence profile are expected to be mutually more similar. The observed correlations confirm that the Align2D procedure introduces additional information in the alignments, specific for the different template structures.

Figure 1A displays the cumulative percentage of the annotated portion of the database that obtained a certain Z-score, depending on which receptor structure was used as a template. In addition, the percentages of annotated sequences that obtained a certain mean Z-score, averaged over all four structures, and the lowest Z-score obtained with any of the four structural templates are plotted. A low Z-score for a model is assumed to allow for the reliable assignment of the LBD fold to a sequence (Sippl, 1993). As shown in Figure 1A, the four individual LBD structures assign this fold to a smaller percentage of the sequences, compared with the strategy of selecting the best score obtained over all four templates. Selecting the best Z-score is warranted, as sequences will
produce better models when these are based on protein structures that are closer in terms of sequence identity or function. The mean Z-score curve is, as expected, intermediate to those of the individual templates. Although Modeler 5.0 allows for the use of all four templates simultaneously in generating models, models generated in this fashion had less favorable Z-scores than the best of all four individual models. The average Z-score of the models obtained when using all four templates simultaneously was \(-3.9\), while the average Z-score obtained using any of the four templates independently was \(-5.9\). In this respect, the ProsalII Z-score is sufficiently sensitive to identify small differences in model accuracy. When all four templates are used to derive models for the template sequences themselves, an average Z-score of \(-8.1\) and an average r.m.s.d. from the crystal structure Ca-positions of 1.6 Å are obtained. When the best model obtained with any of the four templates is selected, these numbers are \(-9.4\) and 0.16 Å, respectively. The fact that using multiple structures simultaneously for homology modeling yields models of lower quality than selecting the best model derived from individual templates is thought to be due to the structural diversity and low sequence identity among the nuclear hormone receptor LBD structures.

Interestingly, the RAR LBD assigns the most favorable Z-scores to the largest number of modeled sequences compared to the other three templates (cf. Figure 1B). This effect is independent of the annotation status of the sequences involved. One possible explanation for this phenomenon is that the RAR LBD crystal structure already has the lowest Z-score (\(-10.1\)) of all templates. The other three LBD structures have Z-scores in the range \(-7.2\) to \(-8.4\). Models based on a high-quality structure could be expected to have an advantage over sequences modeled on templates of lower quality. None of the four LBD structures used in this study were part of the ProsalII 3.0 database of folds that was used to derive the knowledge-based potentials and therefore no intrinsic bias for a particular fold is expected. There appears to be no correlation between the resolution of the various templates and their usefulness in fold assignment. Nevertheless, the four templates appear to be selective for related sequences according to which template yields the best Z-score for the various annotated sequences. For instance, the ER LBD structure provides the best template for estrogen receptor subtype sequences as well as PPAR and DAX1. The PR LBD selects the progesterone receptor sequence as well as the androgen, mineralocorticoid, glucocorticoid and thyroid receptor LBD sequences. The RXR LBD structure yields the lowest Z-scores for all RXR receptor subtypes and a similar trend is observed for the RAR LBD. Clearly, the template structures display a preference for functionally or structurally related sequences. Since gene duplication has been observed in *Caenorhabditis elegans* (Sluder et al., 1999), the original gene may have been a homologue or predecessor of the RXR.

Figure 1C shows a plot of the best Z-scores obtained using any of all four templates versus the cumulative percentage of all models that obtained such a Z-score. The distributions for sequences with annotations in the SwissProt database (Bairoch and Apweiler, 1996) and sequences without annotations are shown in the same figure. Models for unannotated *Caenorhabditis elegans* sequences have less favorable Z-scores than models based on annotated mammalian sequences. This reflects the evolutionary distance between the *Caenorhabditis elegans* sequences and the mammalian template structures used in this study.

Comparison of the two curves in Figure 1C allows for the determination of a threshold below which sequences can be assumed to be reliably annotated to the nuclear receptor LBD domain. Figure 1C shows that not all annotated sequences obtain a low Z-score. The generation of poor quality models for this fraction of the database is due to misalignments caused by low sequence identity and the absence of representative crystal structures for these sequences. Nevertheless, if one assumes that 90% of the annotated sequences can be modeled correctly, a Z-score threshold of approximately \(-4.5\) can be applied to the sequences without annotation. This would provide annotations for \(-70\%\) of the predicted *Caenorhabditis elegans* proteins. The Z-score threshold of \(-4.5\) used here compares well with tests that involved modeling the sequences of the four template structures and two unrelated sequences with different 3D folds.
[taken from PDB entries 1B6U(A) and 1BB9] on all four LBD structures. The resulting models based on unrelated sequences had Z-scores higher than −3 and all models derived from the LBD sequences scored below −6.

When PSI-BLAST is applied to the same database, using the four template sequences individually as starting points, 192 unique sequences are eventually aligned with E-values below 10−4. Of these 192 sequences, 126 were also found using our model building approach. This confirms that sequence alignment protocols and approaches involving homology modeling contribute different information to the process of gene annotation and identify slightly different sets of sequences. Although the somewhat arbitrary choice of threshold in both methods makes a direct comparison difficult, PSI-BLAST performs comparably to our fold recognition approach in terms of the total number of sequences annotated. Interestingly, whereas with PSI-BLAST the RXR sequence identifies the largest number of potential ligand-binding domains (190 sequences compared with 108 for the RAR), the RAR structure allows for the annotation of the largest number of sequences on the basis of homology models (Figure 1B). These differences indicate that sequence annotation through homology modeling is not dominated by sequence homology but actually introduces new, structure-derived information in the annotation process.

Although there is no absolute Z-score criterion defining a correctly annotated sequence, the fact that this analysis yields a sliding scale is not necessarily detrimental. A sliding scale allows the user to select the highest-ranking genes first to be added to a sequence profile or hidden Markov model. Additional sequences could be added in order of their Z-score ranking. As these sequences are selected on structural criteria rather than sequence similarity, they can add new information to the profile or hidden Markov model that cannot be obtained from sequence similarity analysis alone.

Among the top-ranking models for C.elegans predicted proteins, a number of sequences are found without previous annotation for an LBD fold by hidden Markov models in the Pfam database (Bateman et al., 1999). The three highest-ranking sequences are the following Pfam entries (Z-scores, structural template and sequence identities): C56E10.4 (~7.15, RAR, 16.8%), C06B8.1 (~6.90, ER, 12.2%) and ZK418.1 (~6.10, RAR, 17.7%). One model with less than 10% sequence identity with the template structure (Pfam entry T07C5.4, based on the RAR LBD with 9.6% identity) was still assigned a Z-score of ~5.2. The addition of such sequences could broaden the scope of LBD sequence profiles and the Pfam hidden Markov models further.

In conclusion, large-scale homology modeling of nuclear receptor LBDs in combination with model assessment tools provides a useful addition to current gene annotation strategies. The resulting 3D structure-based annotations can be used to refine sequence-based approaches further. The application of homology modeling approaches depends heavily, however, on the availability of high-quality, representative protein structures and sufficient sequence identity between genes and templates. When structures of diverse members of a protein family are available it is advisable to use all of them individually as templates for comparative modeling, especially when mutual sequence identities are low. The preference of a template structure for certain genes may hint at a similar structure and function.

Acknowledgement

The authors thank Dr J.Vervoort of the Laboratory of Biochemistry of Wageningen Agricultural University for the use of the Silicon Graphics cluster for the calculation of ProsaII scores.

References


Received November 22, 1999; revised April 7, 2000; accepted April 17, 2000

394