PROTEIN SUPERFAMILY MEMBERS AS TARGETS FOR COMPUTER MODELING: THE CARBOHYDRATE RECOGNITION DOMAIN OF A MACROPHAGE LECTIN

RONALD E. STENKAMP[1,3], ALEJANDRO ARUFFO[1,2] & JÜRGEN BAJORATH[1,2]

[1]Department of Biological Structure, University of Washington, Seattle, WA 98195

[2]Bristol-Myers Squibb Pharmaceutical Research Institute

3005 First Avenue, Seattle, WA 98121

[3]Corresponding author, phone , fax, electronic mail:

phone: (206)-685-1721

fax: (206)-543-1524

e-mail: stenkamp@u.washington.edu

Abstract

Members of protein superfamilies display similar folds, but share only limited sequence identity, often 25% or less. Thus, it is not straightforward to apply standard homology modeling methods to construct reliable three-dimensional models of such proteins. A three-dimensional model of the carbohydrate recognition domain of the rat macrophage lectin, a member of the calcium-dependent (C-type) lectin superfamily, has been generated to illustrate how information provided by comparison of X-ray structures and sequence-structure alignments can aid in comparative modeling when primary sequence similarities are low.

Introduction

Members of emerging protein superfamilies are generally thought to adopt a similar global fold, despite sharing only low sequence similarity. As targets for comparative protein modeling [1], members of protein superfamilies, for which at least some structural information is available, appear attractive as it is usually possible to generate approximate three-dimensional models for many of these molecules. However, the low level of sequence similarity shared by these proteins makes it difficult to generate topologically meaningful alignments relative to potential structural template(s) [1], and hence the accuracy of such models is often limited and insufficient for more detailed applications.

The well studied immunoglobulin superfamily (IgSF) [2] can be considered a paradigm for the similarities and relationships of sequences and structures in protein superfamilies. Structural studies on a variety of proteins or protein domains belonging to the IgSF, whose sequence similarity is often limited to two conserved cysteines and a few hydrophobic (core) residues, have revealed many variations of the basic immunoglobulin fold [3]. Despite the knowledge of these structures and many IgSF sequences, it remains difficult to identify, by sequence comparison, closely related templates for model building of IgSF molecules with unknown structure.

The situation is even more problematic for protein superfamilies for which, in contrast to the IgSF, only one structure has been determined and where, in consequence, information from structure comparison (ie, the identification of

structurally conserved and variable regions in different proteins) is not available to complement multiple sequence alignments. Until recently, this has been the case for the calcium-dependent (C-type) lectin superfamily [4]. C-type lectin domains, which typically share less than 30% sequence identity, specifically bind mono- or oligosaccharides in a calcium-dependent fashion and function as carbohydrate recognition modules in many mammalian proteins [5]. The C-type lectin domain of the mannose binding protein from rat (MBP) was the first structure of a C-type lectin domain determined [6] and revealed a novel and rather unusual protein fold, consisting of about 50% loops and other extended regions of unusual secondary structure [6]. This structure has served as template for model building of other C-type lectins [7].

More recently, the structure of the C-type lectin domain of E-selectin, a cell adhesion molecule, has been determined [8] as well as structures of the closely related human homolog of rat MBP [9,10]. Structure comparison of MBP and E-selectin [8,11] has, in conjunction with multiple sequence comparison [12], improved the ability to construct meaningful models of other members of the C-type lectin superfamily [11,13]. We illustrate this here by building a model of the C-terminal extracellular carbohydrate recognition domain of the rat macrophage lectin (ML) [14], a type II transmembrane glycoprotein receptor with specificity for galactose and N-acetylgalactosamine [14,15].

Methods

X-ray structures of E-selectin at 2.0 Å resolution [8] and of MBP at 1.7 Å resolution [16] were compared by backbone superposition as described [11]. Structurally conserved regions were identified by sequential least squares superposition of backbone segments of increasing length followed by root mean square deviation (rmsd) comparison. Backbone segments which superimposed with an rmsd of less than 1 Å were determined, and a structure based sequence alignment was generated. This structure-oriented alignment was complemented by multiple sequence comparison of ML, selectins [12] and other C-type lectins (not shown) to better understand the structural relevance of C-type lectin consensus residues. Backbone regions which are structurally conserved in MBP and E-selectin provided the framework for the ML model. The backbone conformations of

structurally variable regions in MBP and E-selectin and regions including insertions and deletions relative to ML were approximated by conformational search [17]. Side chain replacements were modeled in conformations as similar as possible to the original conformation or, for structurally unconstrained positions, using a rotamer search procedure [18]. The initially assembled model was refined by some energy minimization, and the stereochemistry of the refined model and its sequence-structure compatibility [19] were assessed.

Results and Discussion

The sequences of the carbohydrate recognition domains of ML, MBP, and E-selectin are less than 30% identical. Comparison of the MBP and E-selectin X-ray structures reveals significant differences in some regions, although both proteins adopt the same global fold. However, structural conservation in MBP and E-selectin includes two regions of extended unusual secondary structure. The information obtained from structure comparison has been incorporated in a topological alignment of MBP and E-selectin which reflects the spatial equivalence of residues (Figure 1). The comparison allows a better definition of conserved C-type lectin core regions and of variable structural elements (Figure 2) than has been possible on the basis of only the MBP structure. This was illustrated by comparison of E-selectin model and X-ray structures [13]. The knowledge of similarities and differences in these two structures with limited sequence similarity provides the basis for more accurate sequence alignments to model other members of this protein superfamily.

The sequence of the macrophage lectin was aligned against this template by matching core residues, structurally constrained positions and consensus residues in structurally conserved regions. Using these criteria, regions including insertions and deletions could be assigned with some confidence. Structurally conserved regions in MBP and E-selectin were thought to be also conserved in ML and included in the model. Other backbone segments were considered variable, and their conformations were approximated by conformational search calculations. Following side chain modeling, the model was energy refined and its stereochemistry was assessed.

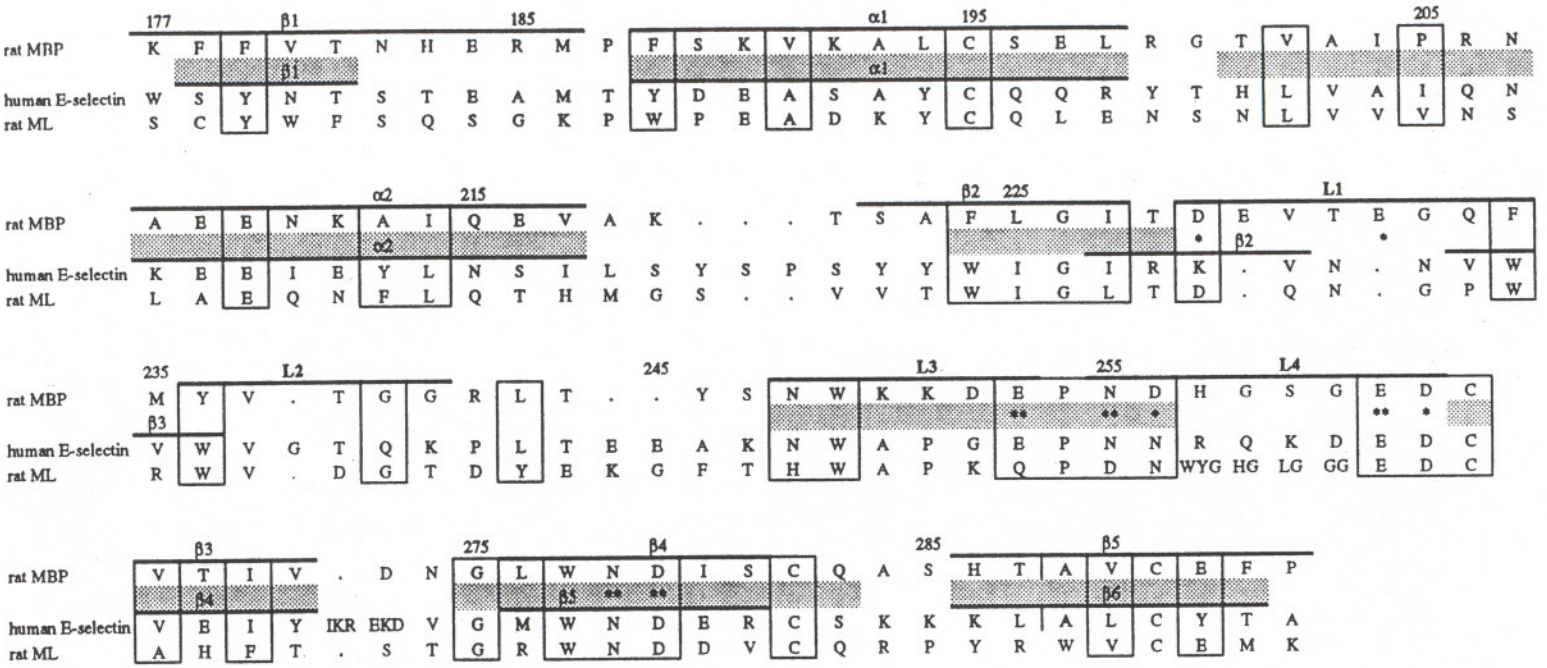

Figure 1. Topological alignment of the C-type lectin domains of human E-selectin and rat MBP, based on structural comparison. The sequence of rat ML was aligned against this template. The major secondary structure elements in MBP and E-selectin are labeled. L1-L4 are extended regions of non-classical secondary structure in MBP. Residues which participate in the formation of two calcium binding sites in MBP are labeled with one and two asterisks, respectively. Residues thought to be determinants the C-type lectin fold are boxed. Horizontal shading shows the structurally conserved regions in MBP and E-selectin. Residue numbers are given for ML.

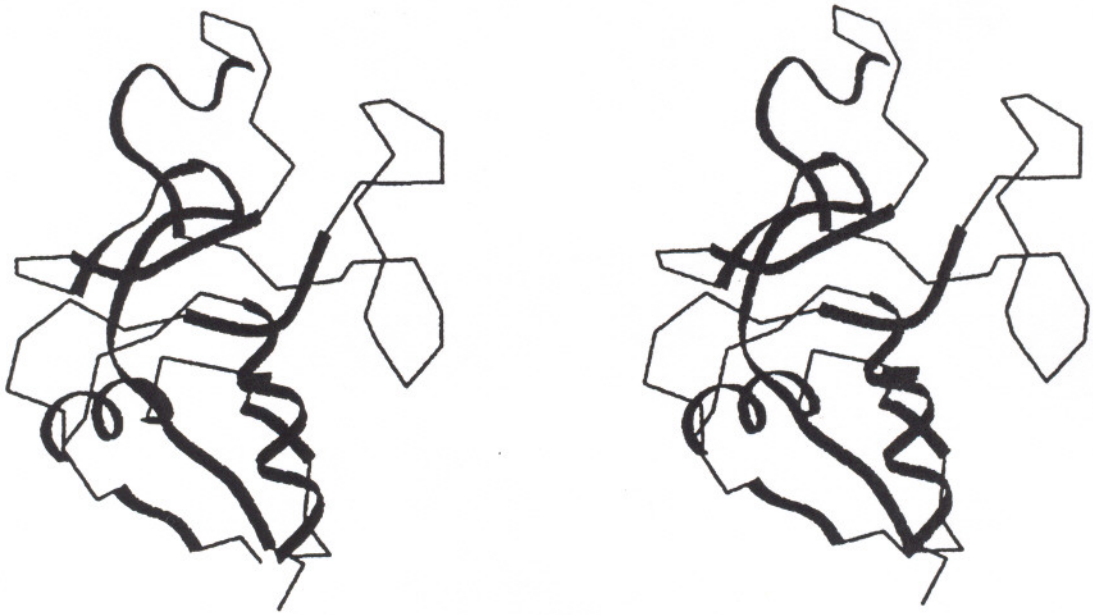

Figure 2. Stereo view of structurally conserved regions in MBP and E-selectin (thick lines) according to Figure 1 mapped on a trace of the MBP C-type lectin fold.

How sound were the assumptions on which the modeling of ML was based? Ultimately, the question can not be answered prior to experimental structure determination. However, the inverse folding approach [20] has provided a variety of methods for threading and the assessment of sequence-structure compatibility which are particularly relevant for model building attempts in the presence of low sequence homology. Here the energy profile method of Sippl [19] has been applied to analyze the sequence-structure compatibility of the ML model (Figure 3). The obtained profile is, in terms of the overall negative value of the average residue interaction energies, indicative of an overall correctly folded model with no substantial errors in the core regions. The analysis is not expected to identify a number of errors such as incorrectly modeled loop conformations, but provides a critical assessment of the reliability of the initial structure-oriented sequence alignment and of the finalized model beyond a determination of its stereochemistry.

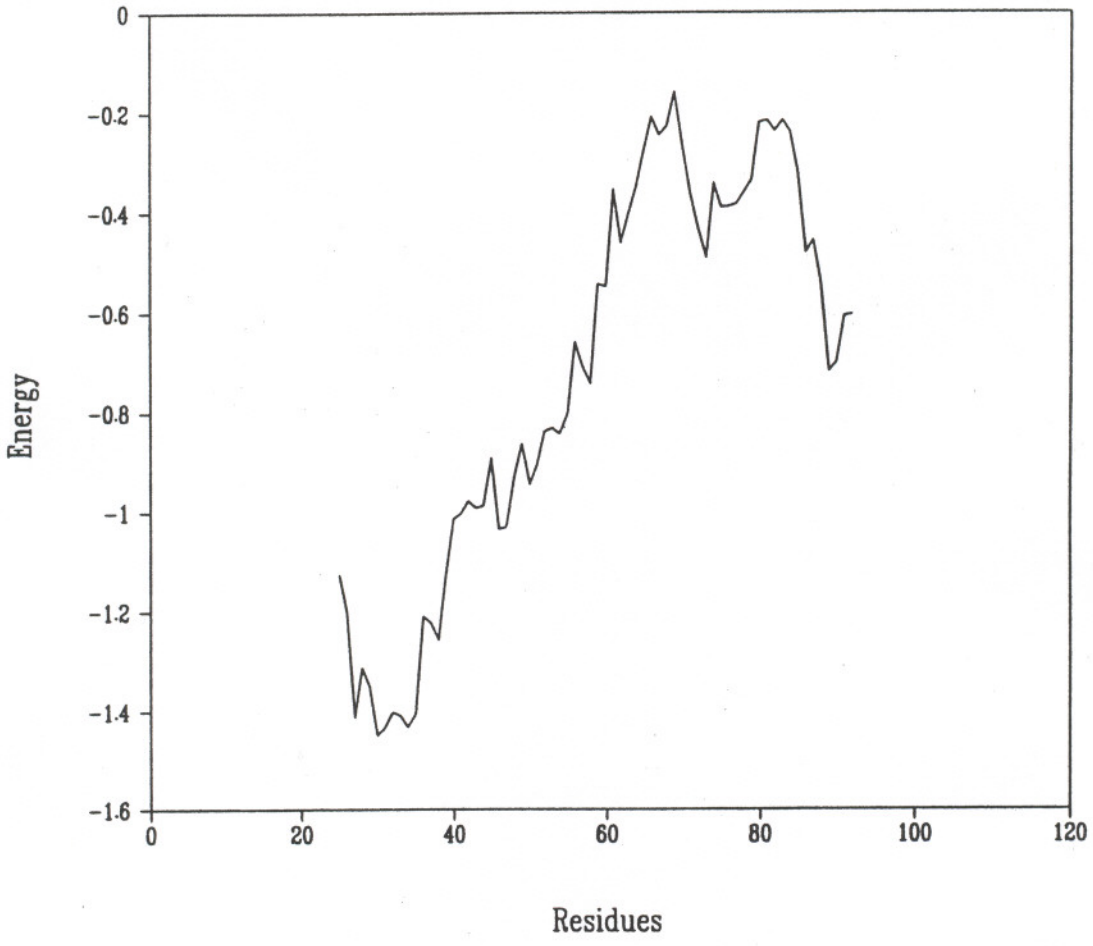

Figure 3. Energy profile of the ML model. Pairwise average residue interaction energy is given in units of E/kT (E, interaction energy in kcal/mol; k, Boltzmann constant; T, temperature in K) and calculated at each residue position using a fifty residue window for energy averaging.

Figure 4 shows the ML model in comparison to the MBP and E-selectin structures. A functional calcium binding site in MBP which is directly involved in carbohydrate binding [16] is predicted to be conserved in ML suggesting that details of the protein-carbohydrate interactions are similar. In contrast, residues surrounding this calcium binding sites and their spatial arrangements differ in all three structures. These ML residues can be selected based on the model and their contribution to specificity can be assessed experimentally. Another unique feature of ML compared to MBP and E-selectin is a five residue insertion in a calcium binding loop.

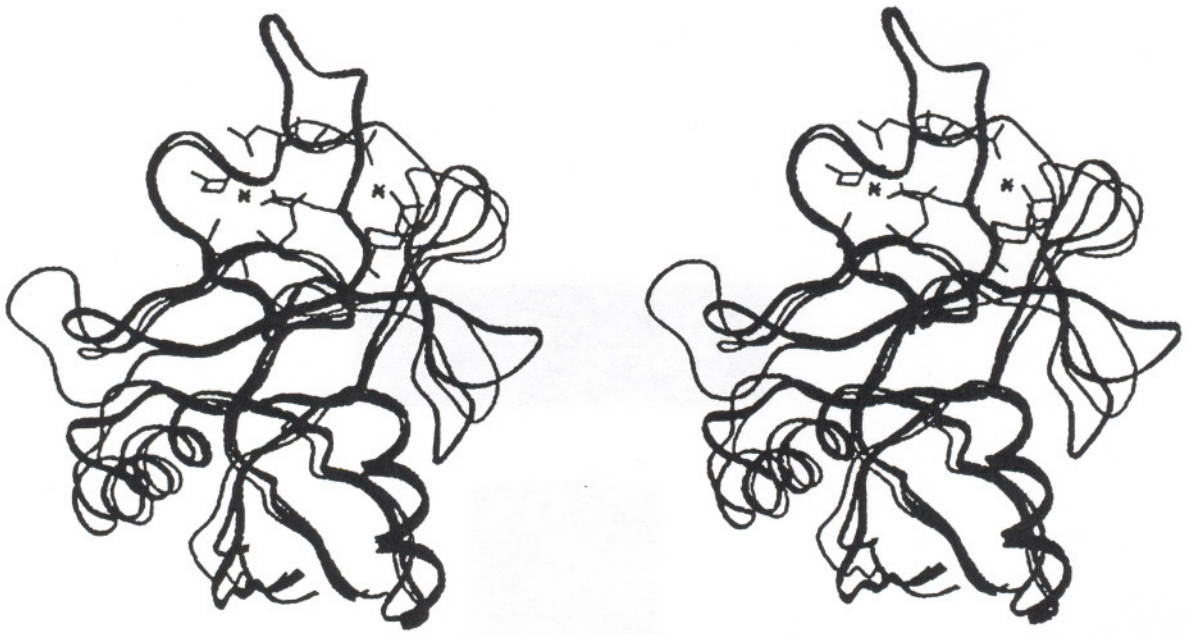

Figure 4. Superposition of C-type lectin domains. The backbones of the MBP and E-selectin structures and the ML model are traced in thin, medium, and thick line, respectively. The orientation is similar to figure 2. ML residues of calcium binding sites fully conserved in MBP, E-selectin, and ML (left) or partially conserved in MBP and ML (right) are shown.

Conclusions

Although members of protein superfamilies are, in principle, promising targets for comparative molecular modeling, the generation of detailed models is often difficult due to low sequence similarities and limited availability of structural data. The generation of reliable models is more straightforward once results from multiple sequence and structure comparison are combined in a meaningful way, as exemplified by studies on C-type lectins such as ML.

References

1. J. Bajorath *et al, Protein Sci.* **2**, 317 (1993).
2. A.F. Williams *et al, Immunol.* **6**, 381 (1988).
3. P. Bork *et al, J. Mol. Biol.* **242**, 309 (1994).
4. K. Drickamer, *J. Biol. Chem.* **263**, 9557 (1988).
5. K. Drickamer, *Curr. Opin. Struct. Biol.* **3**, 393 (1993).
6. W.I. Weis *et al, Science* **254**, 1608 (1991).
7. J. Bajorath and A. Aruffo, *J. Biol. Chem.* **269**, 32457 (1994).
8. B.J. Graves *et al, Nature* **367**, 532 (1994).
9. S. Sheriff *et al, Nature Struct. Biol.* **1**, 789 (1994).
10. W.I. Weis and K. Drickamer, *Structure* **2**, 1227 (1994).
11. J. Bajorath and A. Aruffo, *Protein Sci.* **5**, 240 (1996).
12. J. Bajorath and A. Aruffo, *Biochem. Biophys. Res. Comm.* **216**, 1018 (1995).
13. J. Bajorath et *al, Bioconjug. Chem.* **6**, 3 (1995).
14. M. Li *et al, J. Biol. Chem.* **265**, 11295 (1990).
15. S.T. Iobst and K. Drickamer, *J. Biol. Chem.* **271**, 6686 (1996).
16. W.I. Weis *et al, Nature* **360**, 127 (1992).
17. R.E. Bruccoleri and J. Novotny, *Immunomethods.* **1**, 96 (1992).
18. J. Bajorath and R.M. Fine, *Immunomethods.* **1**, 137 (1992).
19. M.J. Sippl, *Proteins* **17**, 355 (1993).
20. S.J. Wodak and M.J. Rooman, *Curr. Opin. Struct. Biol.* **3**, 247 (1993).