CHAPTER 2

SEQUENCE ALIGNMENT AND KNOWLEDGE-BASED MODEL BUILDING OF GAP
2.1 Introduction

It was predicted that GAP consists of two amphipathic α-helices separated by an intervening loop region (Gupta and Salunke, 1992). Consequently, probable homology with a family of developmental gene regulators that control transcription via a helix-loop-helix (HLH) motif was suggested. It was imperative therefore, to align the sequence of GAP with proteins sharing a common sequence pattern by matching the signature residues. Structural characteristics derived from the sequence alignment were exploited for the construction of a three dimensional (3-D) model for GAP. This chapter details the sequence alignment of GAP and inferences drawn from the pattern in the sequence derived from alignment. Thus, the similarity exhibited could be extrapolated to an equivalence in the mode of action via the formation of a heterodimeric complex of GAP with an ubiquitous factor. A knowledge-based model of GAP, built using the knowledge of the structural similarity with the HLH family of proteins is also discussed in this chapter.

2.2 Knowledge-Based Modeling: Principles & Practice

Extensive information on the primary and secondary structure of proteins is stored in various databases. The advent of recombinant DNA techniques resulted in a faster rate of primary sequence determination as compared to the structures solved for the proteins. Technical problems related to the 3-D structure elucidation by X-ray crystallography or NMR techniques, necessitates the utilization of theoretical procedures for structure prediction. The non-availability of a general rule for folding of a protein, thus relies on ab initio procedures for tertiary structure prediction, or by
Sequence Alignment and Knowledge-Based Model Building of GAP

homology modeling derived from comparing conformations available in the Brookhaven Protein Data Bank (PDB).

Tertiary structures of homologous proteins were observed to possess well conserved 3-D structures retained during the course of evolution. A common core of residues comprising the skeleton of the protein, and the main elements of secondary structure remain highly conserved within a family of homologous proteins. Knowledge-based approach also called homology modeling, is the method of choice for deriving the structure of an unknown protein when the crystal structure of a reference protein is available. The homology modeling approach involves a series of steps which are described in a flow chart in Figure 2.1.

2.2.1 Sequence Alignment

The first step in homology modeling is assignment of the unknown protein to a protein family by searching the databases to find related sequences. Automatic alignment methods were developed to overcome the technical difficulties arising with larger number of proteins for analysis (Gilson et al., 1987; Antosiewicz et al., 1994). Conserved regions identified in the aligned proteins are utilized for transferring the coordinates of the reference protein to the model protein. Initial attempts by Needleman and Wunsch (1970) clarifying the structural similarity between proteins to be dependent on sequence homology led to the subsequent development of alignment programs. An alignment algorithm, a scoring matrix, and a gap-weighting function constitutes a system for optimal alignment of two or more sequences.
Fig. 2.1: Steps followed in Knowledge–based Modeling Approach

Protein Sequence

Internal Pattern Recognition

Structural Module Search

Homology Search
Sequence Alignment
Hydropathy Analysis
2° Structure Prediction

Recognition of Structural Class of Homologous Proteins

Identification of Structurally Conserved and Variable Regions (SCR & SVR)

Modeling SCR Based on Structural Classification

Crude Model Assembly of SCR & SVR

Model Refinement

Structural Module Search for VR Library
In the absence of large number of sequences with adequate homology, search is carried out for well-defined sub-structural modules in the PDB. Common conformational features in different proteins, albeit functionally distinct, are reflected in their polypeptide sequences (Efimov, 1991; Salunke, 1991). Pattern search is therefore, done to discern the common conformational features that can be applied to the unknown. The sequence signatures thus identified from the secondary structural patterns, are applied for deriving knowledge-based models.

2.2.2 Construction of the Model

2.2.2.1 Determination and Generation of Structurally Conserved Regions (SCRs)

The central dogma in homology modeling is that there are certain identical regions in the 3-D structures in proteins belonging to the same family, despite evolutionary changes. These Structurally Conserved Regions (SCRs) are secondary structural units of homologous proteins that occupy the same relative orientation throughout a protein family. Structures in the PDB are superimposed relative to each other, to discover areas that are conserved in specific regions, as well as their overall 3-D structure. The SCRs remain conserved with respect to structure and occasionally the sequence, and do not possess any insertions or deletions. Matching points located in the conserved secondary structural elements and geometrical features are optimized. Thus, coordinates are assigned after correspondences are established, from the reference protein.
2.2.2.2 Construction of Structurally Variable Regions (SVRs)

Significant variations exist in protein structures preferably in the loop regions, which mostly constitute the SVRs. Differences in the number of amino acids result in insertions and deletions. A variety of methods have been developed for the generation of loops (Šali et al., 1990; Jones and Thirup, 1986). The loop search protocol analyzes peptide segments in proteins for their geometrical equivalence, usually by scanning the PDB. Sequence homology as well as the goodness of fit between the original fragment and the retrieved segment is used to rank the loops. De novo generation technique is an alternative method where the peptide backbone chain is built between two conserved segments (SCRs) using randomly generated numerical values for all the backbone dihedral angles. The loop conformations are finally relaxed by energy minimization techniques to remove steric hindrance.

2.2.2.3 Side Chain Modeling

Construction of the peptide backbone is the first step, followed by the addition of side chains. A variety of energetically allowed conformations are adopted by the side chains, as they possess one or more degrees of freedom. Nevertheless, a limited number of conformations are assumed out of the large number of possible choices. Rotamer libraries were developed on the basis of statistical evaluations (Ponder and Richards, 1987) and methods have been derived, considering the local environment and other constraints affecting the side chain conformations (Sutcliffe et al., 1987). Side chains are constructed in light of the steric hindrance and overall conformation of the protein modeled.
2.2.3 Refinement of the Model

In the course of generation of protein models, the loop regions and side chains are chosen arbitrarily; the conformations do not correspond to energetically favorable structures. Removing the constraints at the regions of connection between the SCRs and SVRs is also important. Energy minimization and molecular methods relax the local region of the conformational space occupied and produce a more refined structure.

2.2.3.1 Energy Minimization

Geometry optimization is carried out for the molecular structures to find individual energy minimum state. Molecular mechanics considers the atomic composition of a molecule to be a collection of masses interacting with each other via harmonic forces. In the course of a calculation, the total energy minimized with respect to atomic coordinates is as follows:

\[ E_{\text{tot}} = E_{\text{st}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdw}} + E_{\text{elec}} + \ldots \]

where \( E_{\text{tot}} \) is the total energy of the molecule, \( E_{\text{st}} \) is the bond-stretching energy term, \( E_{\text{bend}} \) is the angle-bending energy term, \( E_{\text{tors}} \) is the torsional energy term, \( E_{\text{vdw}} \) is the van der Waals energy term, and \( E_{\text{elec}} \) is the electrostatic energy term.

Force field is a collection of unstrained bond-terms such as bond lengths, angles etc., in addition to the force constants taken into consideration. The equilibrium values of natural bond lengths and angles and the corresponding force constants used in the potential energy function are defined as the force field parameters. Deviation from the standard values results in an increase in the total energy of the molecule. The dielectric constant is an empirical, global, dimensionless scaling factor that reduces the
size of the charge-charge interaction terms in the potential energy calculation for pairs of non-bonded atoms separated by more than three covalent bonds. A distance-dependent dielectric is used in the refinement procedures to avoid artifacts produced by cutoffs, and to qualitatively mimic the solvent screening effects on electrostatic charges (Guenot and Kollman, 1992).

Energy minimization methods can be divided into two classes: the first-derivative techniques like the steepest descent, conjugate gradient and Powell; and the second-derivative methods like Newton-Raphson and related algorithms. The steepest descent minimizer applies numerically calculated first derivative of the energy function to approach the energy minimum. It is generally used for structures far from the minimum. Conjugate gradient method accumulates information about the function from one iteration to the next for computing the new direction vector, thus guiding towards convergence. The powell method is similar, albeit, faster than the conjugate gradient algorithm. The Newton-Raphson second derivative minimizer, in addition to the gradient, uses the curvature of a function to identify the search direction. Its efficiency is increased as convergence is approached. Minimization algorithms when applied, optimize the geometry and find the local minimum on the potential energy surface closest to the initial coordinates.

2.2.3.2 Molecular Dynamics

The final structure found by minimization algorithms represents only one local minimum. Detecting an energetically most favored 3-D structure requires searching the conformational space in a thorough manner. Molecular dynamics simulations is an ideal solution for the problem, which is performed by integrating the classical equations of
motion over a period of time for the molecular system (Karplus and Petsko, 1990). Unlike the minimization procedures, molecular dynamics overcomes energy barriers between different conformations to find minima other than those nearest on the potential energy surface to the original structure. The essence of this technique lies in the numerical integration of Newton’s second law relating to the mass and acceleration of an atom in the system to the gradient of potential energy. All atoms are assigned velocities, corresponding to a specified temperature determined from the mean kinetic energy of the system \( (1/2 \Sigma m<v^2> = 3/2Nk_BT) \), where \( k_B \) is the Boltzmann constant, \( m \) is the mass, \( v \) is the velocity and the summation is over all the atoms). The velocity and position at a new time point are calculated from the initial value. Conformational sampling is enhanced by applying an elevated temperature to the simulation, that helps overcome large energy barriers existing between some conformations, thereby completing the search. High-temperature annealed molecular dynamics, i.e., the simulated annealing method is also used to enable the molecule escape local minima on the potential energy surface. Molecular dynamics thus helps in escaping from the local minima and for optimizing the side chain conformations.

2.2.4 Validation of the Protein Structure

Determination of the quality and reliability of the model built and minimized is very important, which also depends on the reference protein used. The correctness of the model can be tested on various grounds as depicted in Fig. 2.2. Stereochemical quality of the structure modeled is verified by the accuracy of parameters such as bond lengths, bond angles, torsion angles and correctness of the amino acid chirality.
Fig. 2.2: Validation of Protein models: Methods

Quality of Protein Models

Stereochemical Accuracy
- Torsion angles
  -> Mainchain torsion angle distribution, (Ramachandran plots)
  -> Sidechain torsion angle distribution, ($\chi^1 - \chi^2$ plots)
- Planarity of peptide bonds
  -> $\omega$ angle distribution
- Chirality of C$_\alpha$ atoms
  -> $\xi$ angle distribution
- Bond lengths
- Bond angles
- Planarity
  -> Aromatic ring systems and sp$^2$-hybridized end groups

Packing Quality
- Interatomic distances
  -> 'Bump check'
  -> 'Atomic contact quality'
- Secondary structural elements
  -> Location and geometry of secondary structural elements
- Hydrophobicity
  -> Distribution of polar and nonpolar amino acids
- Solvent accessible surface of amino acids
- Unsatisfied buried H-bond donors/acceptors

Folding Reliability
- 3D-comparison model/template structure
  -> RMS deviations between backbone atoms
- 3D-1D profiles
  -> Comparison of environment strings with amino acid sequences
- Knowledge-based potentials
  -> Energy-based comparison
Specific packing interactions within the interior are also assessed by checking the van der Waals contacts and all the inter-atomic distances. The mean values detected in the crystal structure used as template, are compared with actual values in the generated model. Automated programs like PROCHECK (Laskowski et al., 1993) or WHATCHECK (Vriend, 1990) are applied to determine stereochemical correctness and accuracy of the model. The model refinement is also monitored to look for aberrations during molecular dynamics simulation.

2.3 Modeling of GAP

2.3.1 Sequence Motif

Recognition of a sequence motif in GAP was critical in understanding its structure and function. Based on the predicted secondary structure of the human LHRH precursor, it was suggested that GAP has two helical units separated by an intervening non-helical region (Gupta and Salunke, 1992). Both the helical regions were discovered to have a distinct heptad repeat pattern in the sequence such that the first and fourth residue of each heptad is hydrophobic. Secondary structure prediction led to the identification of a distinct structural motif in the sequence of GAP.

This pattern of heptad repeats with characteristic hydrophobic signature residues has been analyzed in proteins known to interact by coiled-coil folding (Cohen and Parry, 1990; Crick, 1953). Therefore, the sequence signature identified for GAP was aligned with members belonging to the family of developmental gene regulators that contain the HLH structural motif. Alignment of GAP sequence with the developmental gene regulators is illustrated in Fig. 2.3. The heptad repeats with the
**Figure 2.3:** Alignment of GAP with the sequences of the helix-loop-helix proteins. Hydrophobic residues forming the heptad repeat are highlighted in the boxes. The residues corresponding to the two amphipathic α-helices and the loop are schematically represented based on the crystal structure of E47 dimer-DNA complex (Ellenberger *et al.*, 1994).
hydrophobic residues in the sequence are highlighted in the alignment. Significant homology of GAP with the family of HLH proteins can be observed, where the GAP sequence aligns from residues 6 to 49 with the sequences of HLH proteins. Definite conservation is observed in the helical segments as compared with the loop region which differs considerably as seen by insertions or gaps in the alignment (Fig. 2.3). The conservation of sequences is observed to be predominant in the heptad repeat in comparison to the rest of the sequence.

2.3.2 Rationale for Model Building of GAP-E47 Complex

Identification of the HLH motif in GAP sequence shared by the developmental gene regulators is indicative of a structural and functional relationship between GAP and the HLH factors. A possible transcriptional function can be inferred from the homology observed with the group. The sequence of GAP lacks the extension of a region of basic residues at the N-terminal of the motif. Thus, it was identified to be homologous with the negative regulators of the HLH family as described in chapter 1. The three groups of the HLH family of proteins are differentiated by the presence or absence of a basic region of residues at the N-terminal extension of the HLH motif. The positive and ubiquitous factors possess the basic region and therefore can bind to DNA, whereas, the negative regulators lack the basic extension and thus form heterodimers. Homology of the sequence of GAP with HLH proteins, thus indicates a similar molecular mechanism for GAP.

The HLH factors have been observed to exist in a stable state only upon interaction with their dimeric partners. HLH domains from different proteins have
varying abilities to form homo- or heterodimers. The differential affinities for dimer association arise from differences in the interaction between the dimerization partners. These proteins do not have a definite 3-D structure as monomers in solution (Anthony-Cahill et al., 1992). Aggregation of HLH has been found to be facilitated at high concentrations. A few of the proteins whose crystal structures were determined were all in dimeric state (Ma et al., 1994).

It was earlier shown that prolactin secretion, predominantly under inhibitory control, is regulated by hypophysiotropic hormones. GAP was proposed to play an important role in inhibiting the prolactin levels (Nikolics et al., 1985). Many cell extracts had been surveyed for proteins with affinity to bind to E-box proteins to identify similar E-box binding proteins in the endocrine cell lineages. Antibodies to E12/E47 recognize the GH3 cells that have been found to contain a functional HLH domain similar to the insulin enhancer factor (Karlsson et al., 1989). In addition, protein complexes similar to E-box binding transcription regulators have been located in the prolactin-secreting endocrine cell lineages of pituitary origin (Aronheim et al., 1991).

A model for GAP was proposed similar to the complexes formed by negative regulators belonging to developmental gene regulators to explain the possible mechanism of action. Therefore, a heterodimeric complex of GAP with the ubiquitous factor, E47 was constructed applying features observed from the sequence pattern determination.
2.3.3 Procedure

The crystal structure of a homodimer of E47 in complex with DNA at 2.4 Å (Ellenberger et al., 1994) was available. A heterodimeric complex of GAP with E47 was built using the HOMOLOGY module ver. 95.0 (Biosym Technologies, USA). The region corresponding to the residues 6 to 49 of GAP were modeled. The first and fourth hydrophobic residues of the heptad repeat pattern were used as a guide for alignment.

The SCRs corresponding to the two amphipathic α-helices in GAP, were aligned with the template structure E47, and gaps were introduced in the loop region to enable simultaneous alignment of all the SCRs. The coordinates of the SCRs identified in the template structure were transferred to the backbone of the sequence of GAP. The loops (SVRs) were modeled by using the loop search option in the HOMOLOGY module. This option performs alignment of the sequence to be built with loops of structures available in the PDB. Ten best matches presented were analyzed for the best fitting loop with similar \( C^a \) distances and comparable homology. A structural region having moderate homology but with a geometry of only 0.7Å rms deviation was taken for the structure of GAP. The side chains were modeled using the HOMOLOGY module which has a specific rotamer library to assign the coordinates for the side chains.

Refinement by energy minimization and molecular dynamics techniques was carried out to relax the constraints on the model to obtain a stable conformation at its energy minimum. Energy minimization was carried out initially by restraining the backbone and removing steric hindrances on the side chains. AMBER forcefield
(Weiner et al., 1984) was used during various refinement procedures. A low dielectric with a distance-dependent dielectric constant was applied as it aids in reducing the imbalance between the medium-range and long-range electrostatic effects. The program DISCOVER ver. 95.0 (Biosym Technologies, USA), was used to carry out minimization and molecular dynamics. Energy minimization was accomplished first by using the steepest descent algorithm for 100 iterations keeping the total molecule fixed and only the side chain atoms relaxed. A further 100 iterations were then carried out by fixing the trace atoms of the helices and relaxing the loop regions. Conjugate gradient minimizer was finally used to minimize the molecule keeping the Cα atoms of the four helices of the total complex fixed and, relaxing the loop regions and the side chain atoms of the entire model. Minimization was carried out till convergence was reached at a rms energy gradient of 0.001 Kcal/mole/Å. The energy minimized model was subjected to molecular dynamics simulation for 100 ps at 300K. Analysis of the molecular dynamics run was done using DECIPHER ver. 95.0 (Biosym Technologies, USA). Various structures were selected from the simulation to search for an energetically favorable conformation of the model.

The model was validated using the program, PROCHECK (Laskowski et al., 1993). Various bond-terms such as the bond lengths, bond angles, torsion angles etc. were analyzed by the program and the stereochemical favorability was assessed using the Ramachandran plot. Deviations of the van der Waals contacts and other bonding parameters were also analyzed in the structure. The heterodimeric model of GAP-E47 complex was constructed and subsequently analyzed for its structural features using the INSIGHTII molecular modeling software.
2.4 Model of GAP-E47 Complex

The sequence alignment derived from the pattern recognition, was used to build a knowledge-based model of the heterodimeric complex of GAP with an ubiquitous factor E47. The structure was modeled using the coordinates of E47, an ubiquitous factor whose crystal structure was available. SCRs of the dimeric complex, i.e., the four amphipathic α-helices were modeled directly by transferring coordinates from the template structure, E47. The variable regions i.e., the loops were searched from the loop library and finally picked from the structures of pdb1prc.ent and pdb1bkm.ent that had an rms deviation lower than 1Å. The model was then subjected to a minimization and molecular dynamics protocol as described in section 2.3.3.

2.4.1 Energy Characteristics

The structure converged after 2000 iterations of conjugate gradient minimization with a potential energy of -83.0 kcal/mole. Molecular dynamics run was carried out for this structure at 300 K. The energy of the total system undergoes variations during molecular dynamics with an initial increase that corresponds to the rise in the temperature from 0 K to 300 K. This is followed by a less variant status during equilibration and simulation at the same temperature, showing the conservation of energy.

The extent of variation in the structure was analyzed by superimposition of the conformations generated during the MD simulation. The fluctuations occurring during the MD run are more apparent at the loop regions than at the dimeric interface. The helical extension of the E47 unit of the dimer, too undergoes major changes due to the
absence of stabilizing forces as compared to the four helix bundle formed by the α-helices. Root mean square variation of the total energy during the last 40 ps was less than 40 kcal/mole. Ten structures were selected at an interval of 4 ps each and were analyzed for structural variations in the loop region during the simulations. The relaxed loop regions during dynamics simulations have been observed to undergo orientational changes to arrive at the most stable conformation.

There was no significant deviation observed in the overall structure of all the trapped models analyzed. The changes observed are depicted in Figure 2.4, where the core region does not seem to be significantly affected whereas, the loop regions were observed to undergo variations. Root mean square deviations on the side-chain atoms of the core were within 1.5 Å and those for the loop regions were 2.5 Å. The lowest energy conformer out of the set of 40 structures trapped, was selected for further minimization by the conjugate gradient algorithm till convergence was reached. The final model of the dimer of GAP-E47 is shown in a stereoscopic representation in Fig. 2.5. The three dimensional conformation of the monomeric form of GAP highlighting its structural features is shown in Fig. 2.6. Energy of the structure selected at convergence after molecular dynamics simulations is -174.14 kcal/mole.

2.4.2. Structural Features of the Model

GAP and E47 dimerize through their hydrophobic residues to form a stable parallel four helix bundle as shown in Figure 2.5. The dimeric interface is constituted by non-polar residues which are buried in the core. Interactions between various hydrophobic core residues are shown in Fig. 2.7. The N-terminal region of E47 involved in DNA binding is disordered as expected, due to the absence of DNA as also.
Figure 2.4: Diagrammatic representation of the variations occurring in the loop region of the model during molecular dynamics simulations. Superimposition of a set of 5 structures trapped during the run is shown in the figure.
Figure 2.5: Stereoscopic representation of GAP–E47 heterodimeric complex. GAP monomeric unit is depicted in red and its partner E47 in brown colour.
Figure 2.6: A stereo view of ribbon (red) superimposed model of GAP as viewed by E47. The orientation of the $\alpha$-helices is highlighted in the figure. Every tenth residue is labeled in the model.
**Figure 2.7:** Stereodiagram showing the interactions between various hydrophobic core residues of the two chains of GAP–E47 complex. The non-polar residues are depicted in green and GAP and E47 in red and brown respectively.
observed in certain other bHLH (Ferre-D’Amare et al., 1994) and leucine zipper proteins (Talanian et al., 1990). The α-helical extension of E47 in the model built with GAP, was restrained by imposing appropriate constraints. The interactions between the different structural units of the molecule with one another were calculated using the program CONTACT (CCP4). Residues with the most significant interactions are tabulated for the four helices belonging to GAP and its dimeric partner E47, in Table 2.1.

The analysis indicates that loop residues do not show any significant contribution to the two interacting subunits (Table 2.1). The loops, therefore, may be considered to interact with the helices in such a way as to give proper juxtaposition of the helices thereby rendering stability to the fold. The broad features of the model remain similar to those observed for the E47 dimer in the crystal structure of E47 homodimer and also are comparable with the overall conformation depicted in the crystal structure of MyoD-DNA complex (Ma et al., 1994).

The energy of interaction calculated for the heterodimer of GAP-E47 was estimated as the energy required to bring the two monomers from infinity to the correct docking position. Interacting energy (ΔE) for GAP-E47 heterodimer was calculated to be -174 kcal/mole and the corresponding estimate for the E47 homodimer, excluding the DNA was -98 kcal/mole. The buried surface areas (H2O probe size: 1.4) calculated using the program GRASP (Nicholls et al., 1991), between different pairs of helices are tabulated in Table 2.2.

Protein structure validation using the program PROCHECK showed that the heterodimeric complex built was well within the stereochemical limitations. Figure 2.8
Table 2.1: Interactions between and within the two chains of GAP−E47 complex with hydrophobic core residues highlighted in both the chains (G & E). Interchain contacts define the strong hydrophobic associations.

<table>
<thead>
<tr>
<th>INTRA CHAIN CONTACTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>G 8 VAL G 33 ASP</td>
</tr>
<tr>
<td>G 11 VAL G 37 ALA</td>
</tr>
<tr>
<td>G 12 GLY G 33 ASP</td>
</tr>
<tr>
<td>G 15 ALA G 33 ASP</td>
</tr>
<tr>
<td>G 15 ALA G 36 GLY</td>
</tr>
<tr>
<td>G 15 ALA G 37 ALA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>INTER CHAIN CONTACTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>G 3 SER E 28 LEU</td>
</tr>
<tr>
<td>G 4 PHE E 28 LEU</td>
</tr>
<tr>
<td>G 7 ILE E 31 LEU</td>
</tr>
<tr>
<td>G 7 ILE E 32 GLN</td>
</tr>
<tr>
<td>G 8 VAL E 31 LEU</td>
</tr>
<tr>
<td>G 10 GLU E 35 VAL</td>
</tr>
<tr>
<td>G 11 VAL E 31 LEU</td>
</tr>
<tr>
<td>G 11 VAL E 34 ALA</td>
</tr>
<tr>
<td>G 11 VAL E 35 VAL</td>
</tr>
<tr>
<td>G 14 LEU E 38 ILE</td>
</tr>
<tr>
<td>G 14 LEU E 39 LEU</td>
</tr>
<tr>
<td>G 31 LEU E 3 ASP</td>
</tr>
<tr>
<td>G 34 LEU E 4 ILE</td>
</tr>
<tr>
<td>G 34 LEU E 7 ALA</td>
</tr>
<tr>
<td>G 34 LEU E 8 PHE</td>
</tr>
<tr>
<td>G 34 LEU E 11 LEU</td>
</tr>
<tr>
<td>G 34 LEU E 31 LEU</td>
</tr>
<tr>
<td>G 35 LYS E 3 ASP</td>
</tr>
<tr>
<td>G 35 LYS E 7 ALA</td>
</tr>
<tr>
<td>G 37 ALA E 11 LEU</td>
</tr>
<tr>
<td>G 37 ALA E 38 ILE</td>
</tr>
<tr>
<td>G 38 LEU E 10 GLU</td>
</tr>
<tr>
<td>G 38 LEU E 14 MET</td>
</tr>
<tr>
<td>G 40 SER E 38 ILE</td>
</tr>
<tr>
<td>G 41 LEU E 11 LEU</td>
</tr>
<tr>
<td>G 41 LEU E 14 MET</td>
</tr>
<tr>
<td>G 41 LEU E 15 CYS</td>
</tr>
<tr>
<td>G 41 LEU E 18 HIS</td>
</tr>
<tr>
<td>G 41 LEU E 37 VAL</td>
</tr>
<tr>
<td>G 41 LEU E 38 ILE</td>
</tr>
<tr>
<td>G 42 ILE E 14 MET</td>
</tr>
<tr>
<td>G 44 GLU E 38 ILE</td>
</tr>
<tr>
<td>G 44 GLU E 41 LEU</td>
</tr>
</tbody>
</table>
Table 2.2: Buried surface area calculated for various units of the GAP-E47 dimer.

<table>
<thead>
<tr>
<th>Unit of GAP-E47</th>
<th>Surface Area$_{accessible}$ (Å²)</th>
<th>Surface Area$_{buried}$ (Å²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>1550</td>
<td>-</td>
</tr>
<tr>
<td>G2</td>
<td>1664</td>
<td>-</td>
</tr>
<tr>
<td>E1</td>
<td>2025</td>
<td>-</td>
</tr>
<tr>
<td>E2</td>
<td>1714</td>
<td>-</td>
</tr>
<tr>
<td>G1G2</td>
<td>2937</td>
<td>277</td>
</tr>
<tr>
<td>G1E2</td>
<td>2578</td>
<td>686</td>
</tr>
<tr>
<td>G2E1</td>
<td>2931</td>
<td>758</td>
</tr>
</tbody>
</table>
Figure 2.8: Validation of the GAP-E47 complex by PROCHECK program. Ramachandran plot shows that there are no major deviations in the structure.
represents the Ramachandran plot generated for the model. The figure shows that all the residues comply the structural rules and no significant deviations were observed.

2.5 Discussion

Sequence alignment of GAP with the HLH proteins (Fig. 2.3) not only identifies its homology with the group of transcription factors, but also reveals its higher degree of sequence similarity with the negative factors of the group. GAP, similar to other inhibitors of the HLH family, does not possess the additional basic stretch of residues at the N-terminal of the HLH motif. Hence, by corollary, it was conjectured that GAP is functionally similar to the inhibitors and its mechanism of action is equivalent to that of the negative regulators of the HLH family. The structural features of the model built by homology based modeling equate with the features observed in the crystal structure of the homodimer of E47 (Ellenberger et al., 1994). The interactions observed for the model define the hydrophobic residues to be important in maintaining the stability of the dimer. Nonpolar residues of one helix of a subunit associate with the residues in the apposing helices and thereby result in the formation of a core at the dimeric interface (Fig. 2.7). There are also certain hydrophilic residues observed, that seem to play an important role in the overall stability of the structure.

The energy of interaction calculated for the heterodimeric complex of GAP-E47 is comparable to other such protein-protein interactions (Chothia and Janin, 1975). This is indicative of the favorability for the formation of a heterodimeric complex between GAP and an ubiquitous factor. The interaction between GAP and
E47 therefore, is proposed to be a stable interaction favoring the formation of a heterodimer with proper juxtaposition of 4 α-helices such that two α-helices of the same subunit are arranged at an angle of 55°. The opposing helices i.e., helices from different subunits interact at an angle of 15°. Consequently, the surface area buried between two α-helices from the same subunit is significantly lower than that calculated between two helices from apposing subunits. Accessible surface areas calculated for different subunits (Table 2.2) indicate the stability of the heterodimer of the GAP as compared to the interactions within the single unit of GAP.

Identification of HLH domain in GAP therefore, signifies its homology with the developmental gene regulators. The similarity shared, in particular with the negative regulators, enabled the construction of a knowledge-based model of the heterodimeric complex of GAP with E47. The model highlights the favorability for dimer formation and subsequently, of the interactions that are crucial for function. Analysis of accessible surface areas is indicative of the heterodimeric association being critical for maintaining the conformation of GAP. The interactions of the HLH motif in GAP with its putative dimeric partner E47, imply functional relevance to the mechanism of action of GAP. Various structural components of the HLH motif and their associations point out the importance of different units of the motif. Consequently, the sequence alignment of GAP with HLH family of proteins was utilized as a basis for designing peptides that could be tested for their structural characteristics and functional significance. The action of various peptides designed could be correlated with their structural features that would aid in understanding the possible mechanism of action.