Secondary structure prediction is a step towards deducing the fold. In order to arrive at the correct 3D structure, you just need to arrange the secondary structure elements.

Predicted secondary structure can be used to help identify protein function - by searching for similar secondary structural motifs.

Secondary structure prediction can help fold recognition. Many fold recognition methods use a combination of sequence profiles and predicted secondary structure.
Besides,
Sometimes it is All We Have ...

[Diagram showing pie chart with sections for 1D Prediction, Fold Recognition, and Homology Modelling]
Protein Folding is Determined by Amino Acid Sequence ...
Amino acid side chains properties affect packing

Serine: small, polar

Proline: polar, helix breaking

Phenylalanine: large, hydrophobic

Arginine: bulky, polar, basic and hydrophobic
Residues pack into characteristic local structures - alpha helices

The backbone adopts a helical conformation.

Hydrogen bonds between the carboxy group of residue $n$ and the amino group of residue $n+4$.

An ideal alpha helix has 3.6 residues per complete turn.
An ideal alpha helix has 3.6 residues per complete turn.

Helices are flexible

Side chains stick out
Residues pack into characteristic local structures - beta strands

Beta strands form beta pleated sheets because of the stabilising effect of the inter-strand hydrogen bonds.
Side chains stick out above and below the plane of the sheet:

Beta sheets may be parallel, antiparallel or a mixture of both.

Sheets are often twisted or buckled, but are not flexible.
Other local structures

Disorder
Usually not seen in X-ray structures
Mostly polar residues, mobile

Coiled-coil
Long helices

Methods exist for the prediction of all local structure

Loops
Can be “turns”, “bends”, etc.
<table>
<thead>
<tr>
<th>Residue</th>
<th>Helix</th>
<th>β-sheet</th>
<th>Turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu</td>
<td>1.59</td>
<td>0.52</td>
<td>1.01</td>
</tr>
<tr>
<td>Ala</td>
<td>1.41</td>
<td>0.72</td>
<td>0.82</td>
</tr>
<tr>
<td>Leu</td>
<td>1.34</td>
<td>1.22</td>
<td>0.57</td>
</tr>
<tr>
<td>Met</td>
<td>1.30</td>
<td>1.14</td>
<td>0.52</td>
</tr>
<tr>
<td>Gln</td>
<td>1.27</td>
<td>0.98</td>
<td>0.84</td>
</tr>
<tr>
<td>Lys</td>
<td>1.23</td>
<td>0.69</td>
<td>1.07</td>
</tr>
<tr>
<td>Arg</td>
<td>1.21</td>
<td>0.84</td>
<td>0.90</td>
</tr>
<tr>
<td>His</td>
<td>1.01</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td>Val</td>
<td>0.90</td>
<td>1.87</td>
<td>0.41</td>
</tr>
<tr>
<td>Ile</td>
<td>1.09</td>
<td>1.67</td>
<td>0.47</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.74</td>
<td>1.45</td>
<td>0.76</td>
</tr>
<tr>
<td>Cys</td>
<td>0.66</td>
<td>1.40</td>
<td>0.54</td>
</tr>
<tr>
<td>Trp</td>
<td>1.02</td>
<td>1.35</td>
<td>0.65</td>
</tr>
<tr>
<td>Phe</td>
<td>1.16</td>
<td>1.33</td>
<td>0.59</td>
</tr>
<tr>
<td>Thr</td>
<td>0.76</td>
<td>1.17</td>
<td>0.90</td>
</tr>
<tr>
<td>Gly</td>
<td>0.43</td>
<td>0.58</td>
<td>1.77</td>
</tr>
<tr>
<td>Asn</td>
<td>0.76</td>
<td>0.48</td>
<td>1.34</td>
</tr>
<tr>
<td>Pro</td>
<td>0.34</td>
<td>0.31</td>
<td>1.32</td>
</tr>
<tr>
<td>Ser</td>
<td>0.57</td>
<td>0.96</td>
<td>1.22</td>
</tr>
<tr>
<td>Asp</td>
<td>0.99</td>
<td>0.39</td>
<td>1.24</td>
</tr>
</tbody>
</table>

**Strong Helix Forming:** Alanine, Glutamic Acid, Methionine, Leucine

**Strong Strand Forming:** Isoleucine, Valine, Tyrosine, Tryptophan, Leucine

**Strong Turn Forming:** Glycine, Proline
Short-range interactions lead to secondary structure formation.

Secondary structure comes about because of the hydrogen bonds at the local level between the backbone and chain carbonyl and amino groups.

Most prediction methods use neural networks trained on proteins of known structure.

They often also use information from multiple sequence alignments to improve secondary structure prediction.
Secondary Structure Notation

1  ASKGEELFTGVVIPILVELEDGDVDVNGHKFSVSGEGEGDATYGKLTLKFICTT
   TTGGGGSSEEEEEEEEEEEEETTEEEEEEEEEEEEEETTTTTTEEEEEEEEEETT

51  GKLVPWPTRLVTFSYGVQCFSRYPDHMKRHDFFKSAEMPEGYVQERTIFF
    SS SS GGGGHHHHSSSS GGG B GGGGGG HHHHTTTT EEEEEEEE

101  KDDGNYKTRAEVKFEGBTLMNRIELKGDHKDFKDGNILGHKLEYNYNSHV
     TTS EEEEEEEEEETTEEEEEEEEEEEE TTTTTTTT B S EEE

151  YIMADKQKNGIKVFKNFKIRHNIEDGVSQALADHYQQNTPIDGDPVLLPDNHY
     EEEEEGGGETEEEEEEEEEEETTS EEEEEEEEEEEEESSSS SEE

201  LSTQSAKLDPNREKRDHMVLLEFVTAGIT HGMDELYK
     EEEEEEE TT SSSEEEEEEEEEES
Sequence
MREYPVKKGFPTDYDSIKRKISELGFDV
KSEGDLIIASIPGISRIEIKPDKRKILV
NTGDYDSDADKLAVVRTYNDFIEKLTGY
SAKERKKMMMTKD

Prediction Examples

Jpred

SSPro

Prof

SAM T-02
First and Second Generation methods were based on residue propensities and not very reliable.

Predictions for beta sheets were as low as 28-48% and the both helices and sheets were too short.

Beta sheets in particular rely on longer range inter-strand hydrogen bonds for stability.

Stabilising long range interactions were not considered.
Evolutionary information from profiles was critical in improving prediction because evolutionary selection occurs at the structural and functional level and not at the level of sequence.

Conserved and non-conserved patterns from an aligned protein family can be highly indicative of important structural details.

The use of multiple sequence alignments in third generation methods eventually pushed prediction reliability beyond 70%.
Rost and Sander produced a method (PHD) that combined multiple sequence alignments with a neural network designed to bias the underprediction of beta strands.

PHD uses a neural network to achieve a well-balanced prediction of all secondary structure classes.

One important feature is the reliability score.
Multiple sequence information from protein family

Profile derived from multiple alignment for a window of adjacent residues

Two levels of neural network system


Schema for PHD Secondary Structure Prediction
Prediction Reliability Depends on the Protein

Prediction accuracy varies!

Number of protein chains

Per-residue accuracy \((Q_3)\)

\(<Q_3>=72.3\% ; \sigma=10.5\%\)

Prediction Reliability for PHD using a benchmark set of proteins
There are now many methods that use evolutionary information and their reliability is fairly similar. The best methods have reached around 77% reliability, so there has been an incremental improvement in prediction.

Recent improvements are mainly due to larger databases and better multiple alignment methods.

Long-range interactions still not properly considered and there are still occasional confusions between alpha helices and beta strands.

Proteins with unusual characteristics (especially those with few homologues) need to be treated with care.
Secondary Structure Prediction Servers

**Jpred3** - [http://www.compbio.dundee.ac.uk/~www-jpred/](http://www.compbio.dundee.ac.uk/~www-jpred/)
Recent update of Jpred2 that combined the results from four neural networks (JNet, NSSP, Predator, PHD).

**PROFsec** - [http://www.predictprotein.org](http://www.predictprotein.org)
Is based on multiple alignments and other statistics derived from structural databases.

**PSIpred** - [http://bioinf.cs.ucl.ac.uk/psipred/](http://bioinf.cs.ucl.ac.uk/psipred/)
Adds filtered PSIBLAST profiles and neural networks to the results obtained from various secondary structure prediction methods.

**SAM-T02** - [http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html](http://www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html)
A neural network and profiles built using the improved alignments of hidden Markov models.

Uses bi-directional recurrent neural networks to overcome the limitations of feed-forward neural networks with an input window of relative small and fixed size.
There are some regions of sequences that cannot be categorised in one of the secondary structure types.

These regions, generally invisible in crystal structures, are disordered.

Disordered regions are flexible loops, usually characterised by high levels of polar amino acids or low complexity.
Short disordered stretches often found at the start and end of chains but longer loops within chains are mostly conserved in position within families and may have a function.

Possible functions include linkers, spacers, as sites for protein cleavage and in recognition and binding of ligands and other proteins.

They are often found in certain enzymes, such as those involved in cell growth and cell splitting and those involved in protein phosphorylation.

The main enzymes that contain disordered regions are transcription factors, protein kinases and transcription regulators.
One way of assessing these 3D arrangements would be to use predictions of residue solvent accessibility, the extent to which a residue is exposed or buried within the molecule.

If we can reliably predict the secondary structural elements, it might be possible to predict rough 3D structure simply by arranging them in 3D space.
Predicting Solvent Accessibility is Two State

Accessibility is generally predicted by assigning one of two states, buried or exposed, according to residue hydrophobicity.

Sometimes methods introduce degrees of burial, such as 5% exposed, 25% exposed ...
Although accessibility is a function of the hydrophobicity of single residues, simple hydrophobicity analysis is less effective than more advanced methods.

Solvent accessibility prediction can be improved by using residue windows.

Most methods use techniques similar to those used in secondary structure prediction to predict solvent accessibility.
PROFacc - http://www.predictprotein.org
PHDacc and PROFacc employ neural nets and include multiple sequence information. These servers are the only ones that predict real values for relative accessibility (a matrix with values 0, 1, 4, 9, 16, 25, 36, 49, 64, 81).

Jpred - http://www.compbio.dundee.ac.uk/~www-jpred/submit.html
JPred uses PSIBLAST profiles as input for its neural nets and returns the two state possibility "buried/exposed" as its answer.

Accpro - http://scratch.proteomics.ics.uci.edu/
Uses bi-directional recurrent neural networks to overcome the limitations of feed-forward neural networks with an input window of relative small and fixed size.

As you can see the same servers that predict secondary structure predict accessibility.
**Trans-Membrane Proteins**

Known Structures of Transmembrane Protein Domains fall into Two Categories

- α-Helical Bundle
  (Bacteriorhodopsin, PDB 1AP9)

- β-Barrel
  (Matrix Porin, PDB 10PF)
The Trouble with Trans-Membrane Proteins

Trans-membrane proteins are one of the major stumbling blocks in structural genomics.

Reliable computational structure prediction methods are more important since they rarely produce 3D crystals (and are not solvable by NMR).

All atom 3D structural prediction of trans-membrane proteins is still not possible. However, topology prediction is very much possible for helical trans-membrane proteins.
Trans-Membrane Proteins - Almost a 2D Problem

As it turns out, the same strict constraints that hamper crystallisation make the topology prediction of membrane proteins fairly simple. All hydrogen bonds in the lipid-embedded part of the molecule must be satisfied internally, so reducing the degrees of freedom and making prediction almost a 2D problem.

Once the trans-membrane segments are predicted, topology prediction is a matter of exploring all possible conformations of segments.
There are two basic rules governing membrane topology.

The first is that membrane spanning helices tend to be 20-30 residues long and have a high overall hydrophobicity. This makes trans-membrane segments fairly easy to spot from hydrophobicity plots.
Given the location of the transmembrane-segments it is fairly easy to predict membrane helix orientation. Even when some helices are not initially predicted, the rule can help predict the likely topology.

The Positive Inside Rule

This is the second rule.

Loop regions that connect helices on the inside of the membrane (translocated loops) are more positively charged than loop regions on the outside (non-translocated loops).
Topology Prediction Example
Topology prediction methods use a range of hydrophobicity scales and algorithms, and some also use evolutionary information.

Current methods claim that >90% of trans-membrane segments can be correctly identified and that topology can be predicted in >80% of all cases.

However, given the small and biased nature of training sets, some researchers suggest truer figures would be 70% and 60%.
All known membrane beta sheet proteins form beta barrels (porins) that act as passive diffusion pores.

As yet no computational method can predict structure for trans-membrane beta sheet proteins because so few are crystallised.
MEMSAT - http://bioinf.cs.ucl.ac.uk/psipred/
A novel dynamic programming algorithm that makes predictions based on statistical tables compiled from membrane protein data.

TMAP - http://www.mbb.ki.se/tmap/index.html
Uses statistics gleaned from sequence profiles.

PHDhtm - http://www.embl-heidelberg.de/predictprotein/
Combines neural nets, multiple sequence alignments and dynamic programming. The only method with prediction reliability estimates.

TopPred2 - http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html
Averages hydropathy scores with a trapezoid window.
Servers for Trans-Membrane Protein Prediction II

**HMMTOP** - [http://www.enzim.hu/hmmtop/](http://www.enzim.hu/hmmtop/)
The authors define 5 structural states and use hidden Markov models to partition amino acids from the sequence so that the frequency of each amino acid in each state is maximal.

**DAS** - [http://www.enzim.hu/DAS/DAS.html](http://www.enzim.hu/DAS/DAS.html)
Uses multiple alignments from a collection of non-homologous membrane proteins.

**TMHMM** - [http://www.cbs.dtu.dk/services/TMHMM/](http://www.cbs.dtu.dk/services/TMHMM/)
Statistical methods and hidden Markov models help to optimise the localisation and orientation of the transmembrane helices.
Accurate 3D structure prediction will continue to be difficult without sufficient experimental data. However, topology can be determined quickly by combining computational and experimental approaches. Consensus predictions using up to five of the better servers has almost 100% accuracy for TM-helix prediction.
ExPASy Proteomics tools [http://www.expasy.ch/tools/]

PSORT - prediction of signal proteins and localisation sites
TargetP - prediction of subcellular localisation
SignalP - prediction of signal peptides
ChloroP - prediction of chloroplast peptides
NetOGlyc - prediction of O-glycosilation sites in mammalian proteins
Big-PI - prediction of glycosyl-phosphatidyl inositol modification sites
DGPI - prediction of anchor and breakage sites for GPI

NetPhos - prediction of phosphorylation sites (Ser, Thr, Tyr) in eukaryotes
NetPicoRNA - prediction of cleavage sites for proteases in the picornavirus
NMT - prediction of N-miristoylation of N-terminals
Sulfinator - predicts sulphattation sites in tyrosines
Gonzalo Lopez, David de Juan and Amalia Muñoz for various versions of this talk.

Burkhardt Rost, Gunnar von Heijne for the figures I borrowed.