Development and Application of a Free Energy Force Field for All Atom Protein Folding

Abstract

Proteins are the workhorses of all cellular life. They constitute the building blocks and the machinery of all cells and typically function in specific three-dimensional conformations into which each protein folds. Currently over one million protein sequences are known, compared to about 40,000 structures deposited in the Protein Data Bank (the world-wide database of protein structures). Reliable theoretical methods for protein structure prediction could help to reduce the gap between sequence and structural databases and elucidate the biological information in structurally unresolved sequences. In this thesis we explore an approach for protein structure prediction and folding that is based on the Anfinsen's hypothesis that most proteins in their native state are in thermodynamic equilibrium with their environment. We have developed a free energy forcefield (PFF02) that locates the native conformation of many proteins from all structural classes at the global minimum of the free-energy model. We have validated the forcefield against a large decoy set (Rosetta). The average root mean square deviation (RMSD) for the lowest energy structure for the 32 proteins of the decoy set was only 2.14 Å from the experimental conformation. We have successfully implemented and used stochastic optimization methods, such as the basin hopping technique and evolutionary algorithms for all atom protein structure prediction. The evolutionary algorithm performs exceptionally well on large supercomputational architectures, such as BlueGene and MareNostrum. Using the PFF02 forcefield, we were able to fold 13 proteins (12-56 amino acids), which include helix, sheet and mixed secondary structure. On average the predicted structure of these proteins deviated from their experimental conformation by only 2.89 Å RMSD.

Entwicklung und Anwendung eines Kraftfelds für die freie Energie zur Proteinfaltung mit atomarer Auflösung

Zusammenfassung

Proteine sind die nano-skaligen Maschinen der Zelle. Sie sind Bausteine und Funktionseinheiten aller Zellen und funktionieren typischerweise in spezifischen dreidimensionalen Konformationen, die sie als Endpunkt eines komplexen Faltungsprozesses annehmen. Gegenwärtig sind über eine Million Proteinsequenzen bekannt, es konnten jedoch nur etwa 40.000 Strukturen von Proteinen aufgelöst und in der Proteindatenbank hinterlegt werden. Verlässliche theoretische Methoden zu Proteinstrukturvorhersage könnten helfen, diese Lücke zwischen den Sequenz- und den strukturellen Datenbanken zu schließen und die biologische Information in den bislang strukturell unbekannten Proteinen zu entschlüsseln. In dieser Dissertation untersuchten wir einen Ansatz zur Proteinstrukturvorhersage und -faltung, der auf Anfinsons thermodynamischer Hypothese aufbaut, nach der sich Proteine in ihrem nativen Zustand im Gleichgewicht mit ihrer Umgebung befinden. Wir entwickelten daher ein Kraftfeld für die freie Energie von Proteinen (PFF02), das die nativen Konformationen vieler Proteine aller bekannten Strukturklassen als das globale Minimum des Modells der freien Energie beschreibt. Wir haben dieses Kraftfeld gegen die Strukturen des Rosetta Testdatensatzes getestet und fanden, dass die Strukturen mit der jeweils niedrigsten Energie für 32 Proteine dieses Datensatzes im Mittel nur 2,14 Å von der assoziierten experimentellen Konformation abwichen. Wir haben darüber hinaus stochastische Optimierungsverfahren, unter anderem die Basin-Hopping Methode und evolutionären Algorithmen, für die Proteinstrukturvorhersage und -faltung mit atomarer Auflösung entwickelt. Insbesondere der evolutionäre Algorithmus lieferte auf großen Supercomputern, wie zum Beispiel den BlueGene oder MareMonstrum Supercomputer-Clustern, hervorragende Ergebnisse. Mit dem PFF02 Kraftfeld waren wir in der Lage, 13 Proteine mit 12-56 Aminosäuren Länge mit helikaler, Faltblatt- oder gemischter Sekundärstruktur zu falten. Im Mittel wichen dabei die vorhergesagten Strukturen von den jeweiligen experimentell bekannten Strukturen dieser Proteine um nur 2,89 Å RMSD ab.