Eine neue Methode zum Proteinstrukturalignment und ihre Anwendung auf zirkulär permutierte Proteine

Dissertation zur Erlangung des akademischen Grades des

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie

der Freien Universität Berlin

vorgelegt von

Tobias Schmidt-Gönner

aus Mannheim

Februar, 2010

Diese Dissertation wurde finanziert von der DFG über das

IRTG Genomics and Systems Biology of Molecular Networks

1 Gutachter: Prof. Dr. Ernst-Walter Knapp

2. Gutachter: Prof. Dr. Udo Heinemann

Disputation am 26.05.2010

Vorwort

1864 träumte Friedrich August Kekulé von Stradonitz angeblich in einem Tagtraum die Struktur des Benzols, welche die Chemie seiner Zeit vor ein Rätsel stellte. Für die 1838 von Jöns Jakob Berzelius entdeckten Proteine, hatte jedoch bisher noch niemand einen ähnlichen Traum. Die Entschlüsselung der Proteinstruktur ist eine Herausforderung und ein bisher nicht gelöstes Problem.

Kann man die Entschlüsselung der Struktur des Benzols mit der Struktur von Proteinen vergleichen? Kann man das Proteinstrukturproblem als nicht gelöst ansehen? Schließlich wurden Max Perutz und John Kendrew bereit 1962 für ihre Arbeit zur Aufklärung der ersten bekannten Proteinstrukturen mit den Nobelpreis für Chemie ausgezeichnet.[1,2] Seitdem sind mittels Kristallographie und NMR inzwischen viele tausend Proteinstrukturen bekannt. Die Protein Data Bank PDB[3] umfasst über 50000 bekannte Proteinstrukturen.

Es gibt durchaus Ähnlichkeiten zwischen den beiden Problemen. Kekulé wusste, aus welchen atomaren Bausteinen Benzol sich zusammensetzte, wusste um die Vierbindigkeit des Kohlenstoffes, und folgerte daraus die Struktur des Benzols.

Wir wissen heute alles notwenige über die Zusammensetzung der Proteine, ihre Strukturbestandteile und Regeln, aber eine zuverlässige Strukturvorhersage für Proteine ist immer noch nicht greifbar.

Im Rahmen dieser Doktorarbeit habe ich nach diesem Stern der Erkenntnis gegriffen, ihn aber nicht erreicht. Das Ziel einer zuverlässigen Proteinstrukturvorhersage bleibt in Ferne. Aber, ähnlich wie in dem Berufszweig, in dem die Menschheit am offensichtlichsten nach den Sternen greift, der Raumfahrt, ist auf dem Weg dahin einiges an Nützlichem abgefallen. So wie das Teflon, das zur Beschichtung der Raumkapsel dienen sollte den Weg in die Bratpfannen der Welt gefunden hat, so ist aus den Anstrengungen eine große und interessantere Datensammlung an Decoys zu erstellen, die für die Proteinstrukturvorhersage nützlich ist, ein Programm zum Alignment von Proteinstrukturen hervorgegangen. Ein Programm, dass seinen eigenen Nutzen hat, wie er in dieser Arbeit auch dokumentiert ist.
Ich möchte an dieser Stelle all jenen Danken die mich auf dem Weg dorthin begleitet haben, insbesondere meinem Doktorvater Ernst Walter Knapp, der mir die Möglichkeit gegeben hat einem Traum nachzujagen, Björn Kolbeck und Patrick May mit denen zusammen das Strukturalignmentprogramm GANGSTA entwickelt wurde und Aysam Guerler für die

Weiterentwicklung zu GANGSTA+ und die Umsetzung der neuen Alignmentoptionen, sowie dem Graduiertenkolleg Genomics and Systems Biology of Molecular Networks, das einen Teil dieser Arbeit finanziert hat.

Diese Doktorarbeit basiert auf den folgenden Journal Artikeln:

Kolbeck B, May P, Schmidt-Goenner T, Steinke T, Knapp EW.

Connectivity independent protein-structure alignment: a hierarchical approach.

BMC Bioinformatics 2006,7(1):510.

Schmidt-Goenner T, Guerler A , Kolbeck B und Knapp EW.

Circular permuted proteins in the universe of Protein folds

Proteins: Structure, Function, and Bioinformatics 2009, 78(7): 1618

http://www3.interscience.wiley.com/cgi-bin/fulltext/123218318/HTMLSTART

doi:10.1002/prot.22678

Inhaltsverzeichnis

Einleitung

Proteine

Proteine sind die multifunktionellen Bausteine des Lebens. Zellstruktur, Signaltransduktion, Stoffwechsel, Stofftransport und Ladungstransport, für all diese unterschiedlichen Aufgaben gibt es spezielle Proteine. Trotz dieser vielfältigen Aufgaben setzt sich der Großteil aller Proteine aus Ketten aus nur 20 verschiedenen Aminosäuretypen zusammen.

Die Aminosäuren haben alle die gleiche Grundstruktur:

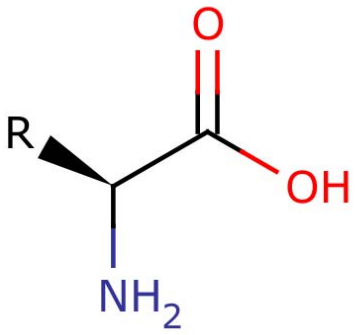
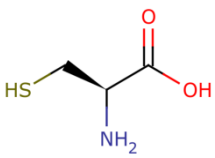
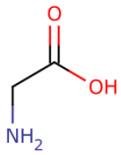


Abbildung 1: Aminosäure Grundstruktur

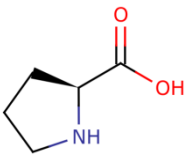
Die Aminosäuren basieren unterscheiden sich nur im Rest R, der Seitenkette.



Cystein



Glycin



Prolin

Abbildung 2: Besonders strukturrelevante Aminosäuren

Der Reihe nach von links nach rechts:

Cystein, welches über Sulfidbrücken die 3D-Struktur maßgeblich beeinflussen kann, Glycin, welches durch die kleinste Seitenkette die größten sterischen Freiheiten bietet und deshalb besonders häufig in engen Schleifen vorkommt und Prolin, welches durch die Ringbildung sterisch am wenigsten flexibel ist.

Zwei aufeinanderfolgende Aminosäuren in der Kette sind über Peptidbindungen miteinander verknüpft. Die Peptidbindung wird durch Elimination von Wasser aus der Carbonsäure der ersten Aminosäure und der Aminogruppe der zweiten Aminosäure gebildet.

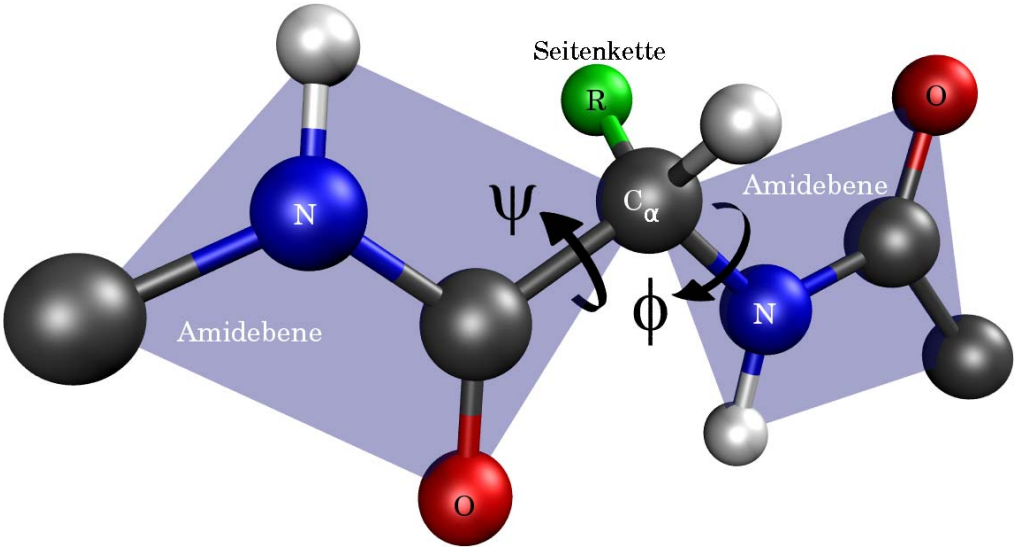


Abbildung 3: Die Peptidbindung

Die Peptidbindung, obwohl keine echte Doppelbindung hat durch die Bildung eines mesomeren Systems einen leichten Doppelbindungscharakter. Dieser sorgt dafür, dass im Proteinrückgrat nur um zwei Bindungen pro Monomereinheit rotiert werden kann. Über die Winkel Ψ und Φ kann daher die gesamte Proteinstruktur beschrieben werden. Die Verteilung der in nativen Proteinen vorkommenden Ψ und Φ -Winkel wird im Ramachandran-Plot dargestellt. Dabei kann man sehr klar erkennen das bestimmte Winkelkombinationen dominieren.

Diese Winkelkombinationen sind typisch für bestimmte Sekundärstrukturen. Unter Sekundärstruktur von Proteinen versteht man spezifische räumliche Anordnungen in einem kurzen Abschnitt der Aminosäurekette. Die zwei am weitesten verbreiteten regulären Sekundärstrukturen sind die α-Helix und das β-Faltblatt, die sich durch eine Wiederholung der selben Φ−Ψ Kombinationen auszeichnen und die durch ein typisches Muster an Wasserstoffbrücken stabilisiert werden. Als drittes Strukturmotiv ist das Random Coil zu nennen, wobei die Einordnung von Random Coil als eigene Sekundärstruktur fragwürdig ist, weil es eben das Fehlen einer festen Struktur darstellt. Weitere bekannte wenn auch häufig in die anderen Klassen eingeordnete Sekundärstrukturmotive wie die 3-10-Helix, die π-Helix und die β-Brücke werden in dieser Arbeit nicht benutzt, da sie nur selten auftreten. Deshalb werden im Folgenden nur die α-Helix und das β-Faltblatt im Detail beschrieben.

Die α-Helix ist eine rechtshändig gedrehte Spirale mit durchschnittlich 3,6 Aminosäureseitenketten pro Umdrehung. Die Ganghöhe beträgt dabei 5,4 Å. Stabilisiert wird sie durch eine Wasserstoffbrückenbindung zwischen dem Carbonylsauerstoff der n-ten und der Aminogruppe der (n+4)-ten Aminosäure desselben Moleküls.

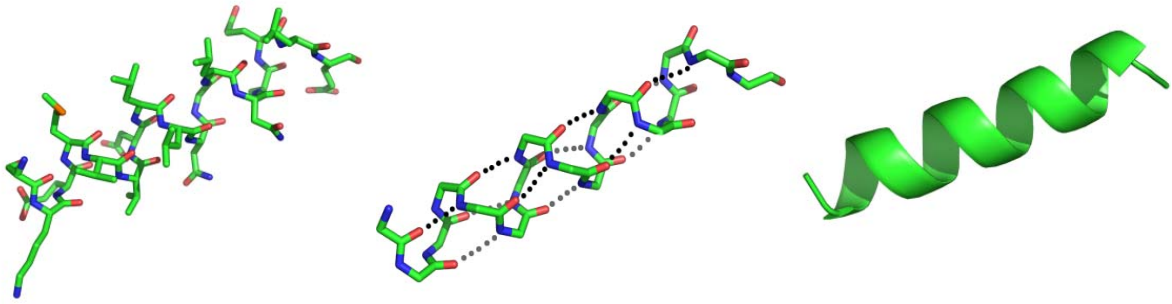


Abbildung 4: Alpha Helix
Von links nach rechts: Vollständiges Stickmodell, Polypeptidrückgrat mit Wasserstoffbrücken und Cartoondarstellung

Bei β-Faltblättern liegen Teile der Aminosäurenkette längs nebeneinander, entweder parallel oder antiparallel. Die stabilisierenden Wasserstoffbrücken kommen in Zweierpaaren im Abstand von 7,0 Å vor. Damit es den einzelnen Aminosäureketten möglich ist, sich dichter nebeneinander zu legen und die zur Stabilisierung notwendigen Wasserstoffbrücken zu bilden, sind die β-Faltblätter sind nicht flach sondern ziehharmonikaähnlich geriffelt, was die Seitenketten aus der Ebene verbannt. Da die Wasserstoffbrücken
nicht wie bei der α-Helix, zwischen innerhalb der Kette nahe beieinanderliegenden Aminosäuren vorliegen, kann ein Faltblattstrang nicht isoliert vorliegen. Es bedarf mindestens zweier Stränge, damit sich ein stabiles Faltblatt bilden kann.

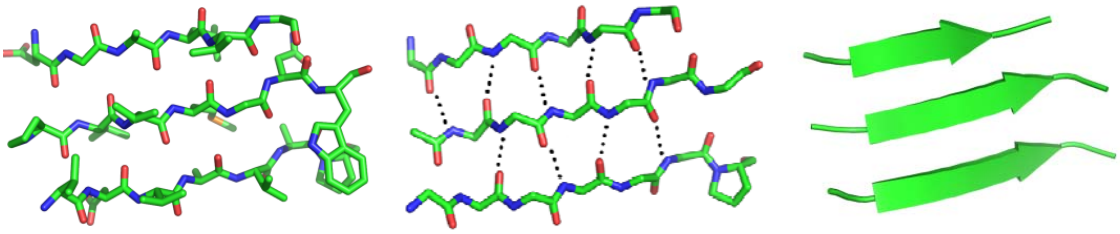


Abbildung 5: Paralleles Beta-Faltblatt
Von links nach rechts: Vollständiges Stickmodell, Polypeptidrückgrat mit Wasserstoffbrücken und Cartoondarstellung

Übergeordnet in der Hierarchie der Proteinstruktur gibt es noch Tertiärstruktur und Quartärstruktur. Unter der Tertiärstruktur versteht man die vollständige dreidimensionale Struktur einer einzelnen Polypeptidkette. Wenn sich mehrere Aminosäureketten zu einem funktionellen Komplex zusammenlagern, so bezeichnet man deren räumliche Anordnung als Quartärstruktur. Während Wasserstoffbrücken von Proteinrückgrat und Seitenketten, ionische Bindungen, und Van-der-Waals-Kräfte sowie der hydrophobe Effekt auch bei der Ausbildung von Sekundär und Tertiärstruktur eine wichtige Rolle spielen, beruht die Ausbildung spezifischer Quartärstrukturen fast nur auf diesen. Nur in sehr seltenen Fällen ist über Sulfidbrücken der Cysteine eine kovalente Bindung an der Quartärstruktur beteiligt.

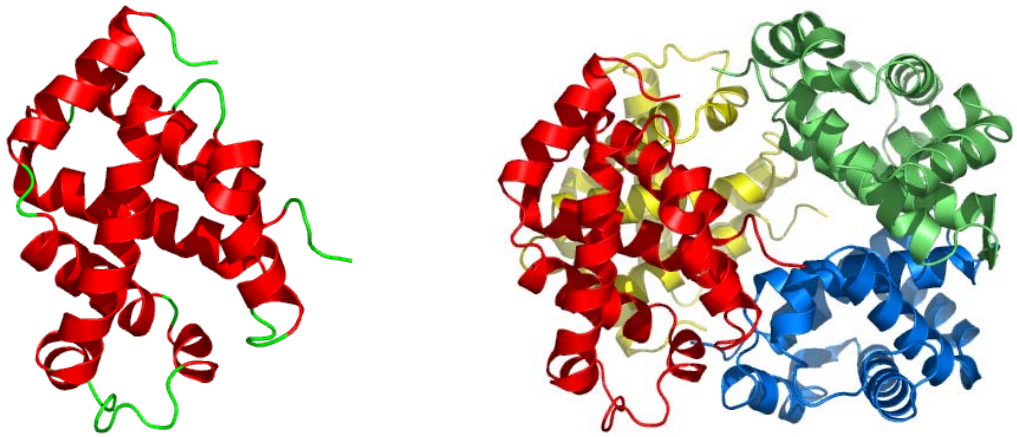


Abbildung 6: Vergleich Tertiärstruktur zu QuartärStruktur
Tertiärstruktur einer Kette des Hämoglobins (links) und Quartärstruktur des Hämoglobins (rechts)

Die Struktur eines Proteins wird in der Regel durch seine Aminosäuresequenz bestimmt.[4] Dies kann durch Rückfaltungsexperimente gezeigt werden, bei denen Proteine durch ändern der Randbedingungen unter denen sie stabil sind vorsichtig denaturiert werden, um dann zu physiologischen Bedingungen zurückzukehren. Dabei falten sich die meisten Proteine von alleine wieder in ihre ursprüngliche Form, was sich z.B. bei Enzymen durch Verlust und Rückgewinnung ihrer katalytischen Aktivität zeigt. Das Rückgewinnen ihrer Aktivität bei Wiedergewinnung der 3D-Struktur zeigt außerdem auf, das die Funktion eines Proteins von seiner Struktur bestimmt wird. Es gilt daher auch: Je ähnlicher sich zwei Proteinstrukturen sind, desto wahrscheinlicher erfüllen sie eine ähnliche Aufgabe. Daher gibt es auch eine ganze Reihe von Programmen für den Strukturvergleich von Proteinen.

Proteinstrukturalignment

Die meisten Proteinalignmentprogramme basieren auf dem Alignment der Sequenzen, die ja schließlich auch die Struktur inhärent bestimmt. Die DNS-Sequenzen bzw. die daraus folgenden Aminosäuresequenzen sind einfacher zu analysieren und auch einfacher zu erhalten als die 3D-Strukturen. Das menschliche Genom z.B. ist durchsequenziert, alle Proteinsequenzen liegen vor. Aber es gibt durchaus gute Gründe weshalb man sich nicht alleine auf das Vergleichen der Sequenz verlassen sollte. Zum einen entstehen die meisten Proteine mit ähnlichen Aufgaben zwar durch Mutationen aus gemeinsamen Vorgängern, dennoch kann auch konvergente Evolution dahin führen. Dabei entstehen Ähnlichkeiten in der Struktur, aufgrund der Ähnlichkeit der Aufgabe, ohne dass es einen gemeinsamen evolutionären Vorgänger gab. Zum anderen sind auch weiter entfernt verwandte Proteine durch ein Strukturalignment zu erkennen, da die Struktur sich weniger schnell ändert als die Sequenz. Die Grenze für erfolgreiches Sequenzalignment liegt in der Nähe von 30% Sequenzidentität. Für die in dieser Arbeit beschriebenen Untersuchungen zirkulär permutierter Proteine wurde beispielsweise ASTRAL40[5] benutzt, eine Untermenge der Strukturen in der PDB[3], die untereinander weniger als 40% Sequenzidentität aufweisen. Viele Proteinpaare aus diesem und ähnlichen Datensätzen sind deshalb mit Sequenzalignment nicht sinnvoll vergleichbar.
Proteinstrukturalignment kann auch für solche Datensätze noch brauchbare Ergebnisse liefern, es gibt also gute Argumente für das Proteinalignment auf Strukturbasis. Allerdings ist das Strukturalignment vom rechnerischen Aufwand auch deutlich höher als der für das sequenzbasierte Alignment. Daher müssen alle strukturbasierten Programme Einschränkungen des Suchraumes vornehmen, um mit der heute vorhandenen Hardware in akzeptabler Zeit zu einem Ergebnis zu kommen. Eine der häufigsten benutzten Einschränkungen ist eine Aufrechthaltung der sequenziellen Reihenfolge, obwohl Proteinstrukturen häufiger auch nicht sequenziell aufeinander abgebildet werden können. Das in dieser Arbeit vorgestellte Programm GANGSTA[6] und sein Nachfolger GANGSTA+[7] benutzen eine besondere hierarchische Strategie um den Suchraum einzuschränken, In einer ersten Stufe wird das Protein auf seine Sekundärstrukturelemente reduziert. Diese Reduktion erlaubt eine Ausweitung des Suchraumes auch auf nicht sequenziell eingeschränkte räumliche Anordnungen. In der zweiten Stufe, wird die grobe Zuordnung aus der ersten Stufe übernommen und dann auf der Ebene der Aminosäuren ausgebaut und vervollständigt. Da nicht immer die beste grobe Zuordnung auch zu der optimalen Lösung des Strukturalignments auf der zweiten Stufe führt, werden aus der ersten Stufe gleich

mehrere mögliche Lösungen in die zweite Stufe des Alignments übernommen und unter ihnen nach der besten Lösung gesucht. Die auf diese Art erhaltenen Ergebnisse sind selbst in Fällen von sequenziell gut aufeinander abbildbaren Proteinen qualitativ mit denen anderer Alignmentprogramme vergleichbar und während GANGSTA gleichzeitig auch für nicht sequenzielles Strukturalignment einsetzbar ist. Eine Anwendung, die ein nicht sequenzielles Strukturalignment erforderlich macht, ist das Strukturalignment von zirkulär permutierten Proteinen.

Zirkulär permutierte Proteine

Proteine bestehen wie weiter oben beschrieben aus Ketten von Aminosäuren die untereinander über Peptidbindungen verknüpft sind. Verbände man die beiden endständigen Aminosäuren eines Proteins ebenfalls über eine Peptidbindung, so erhielte man eine geschlossene Kette. Die Aminosäuresequenz ließe sich dann als ein Kreis, ohne Anfang und Ende darstellen. Wenn man nun diese Kette an anderer Stelle wieder auftrennt, erhält man ein Protein mit anderer Aminosäuresequenz. Ein zum Original zirkulär permutiertes Protein. Seit Jahren sind Beispiele dafür aus der Natur bekannt und auch künstlich im Labor erzeugt worden. In der Natur entstehen zirkulär permutierte Proteine wohl hauptsächlich auf der DNS-Ebene, durch einen Verdopplungsvorgang mit anschließendem Verlust der überflüssigen Teile.

Es werden jedoch auch andere Mechanismen diskutiert und nicht immer findet eine solche Permutation auf Sequenzebene statt. Für Concanavalin A ein Lektin aus der Jackbohne z.B. wurde gezeigt, dass es durch Splicing auf der Peptidebene zu seinem Vorläuferprotein eine Permutation der Aminosäuren 1-118 und 119-237 aufweist[8]. Deshalb erhält man bei einem Alignment mit dem Lektin aus der Erbse oder der Saubohne eine zirkulär permutiertes Proteinpaar.

Vielfach nehmen die zirkulärpermutierten Proteine dabei eine 3D-Struktur ein, die der des Originalproteins extrem ähnelt. Da auch nach ligieren und auftrennen der Kette die gemeinsame 3D-Struktur viele bevorzugte Interaktionen zwischen den Aminosäuren erhält, ist die Struktur des einen Partnerproteins auch logischerweise für das andere Protein eine energetisch sinnvolle Struktur. Aber die direkte Verknüpfung der Aminosäuren zwingt einer Struktur lokal natürlich starke Restriktionen auf, die an der Schnittstelle dann wegfallen, beziehungsweise an der Verknüpfungsstelle dazukommen. Diese lokalen Unterschiede

können dazu führen, dass sich eine komplett andere, vorher eventuell sogar unmögliche Struktur nun als energetisch günstiger herausstellt. Somit lässt sich an zirkulär permutierten Proteinen die Stabilität von 3D-Strukturen besonders gut untersuchen. Daher wurde schon früh eine Möglichkeit gefordert um Zirkulär permutierte Proteine gezielt zu finden.[9]

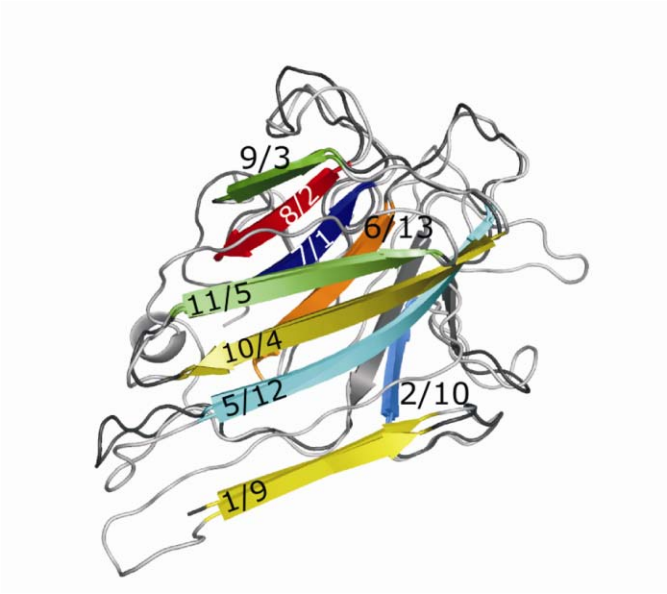


Abbildung 7 GANGSTA Strukturalignment für das zirkulär permutierte Proteinepaar Lektin (1rin) aus der Erbse und Concanavalin A (2cna).

Die Unterbrechung in der Sequenz der Kette führt jedoch dazu, dass mit einem Sequenzalignment solche Proteine nicht mehr ganz so einfach als verwandt erkannt werden können. Der für das Sequenzalignment gebräuchliche Needleman-Wunsch-Algorithmus[10] für das globale Alignment kann unter optimalen Beddingungen den kleineren versetzten Kettenteil nur als Lücke interpretieren. Da Lücken immer zu einem Strafterm führen, wird das globale Alignment dadurch schlechter. Ein lokales Alignment zum Beispiel durch den Smith-Watermann Algorithmus[11] wird dadurch nicht beeinträchtigt. Durch Verschieben der Sequenzen gegeneinander und Bewertung der lokalen Maxima ist es dann möglich zirkulär permutierte Proteine mittels Sequenzalignment zu entdecken.

Gleichzeitig führt die Kettenunterbrechung natürlich auch dazu, dass alle sequenziell arbeitenden, strukturbasierten Alignmentwerkzeuge vor dem selben Problem stehen, da auch hier maximal die größere Hälfte des Proteins sinnvoll aufeinander gelegt werden kann. GANGSTA kann jedoch zirkulär permutierte Proteine beidseits der Schnittstelle zuordnen. Zusätzlich ist es möglich an eben diesem Bruch in der Sequenzreihenfolge der Sekundärstrukturelemente sogar eine Zirkulärpermutation erkennen.

Veröffentlichungen

Connectivity independent protein-structure alignment: a hierarchical approach

Autoren : Björn Kolbeck , Patrick May, Tobias Schmidt-Gönner, Thomas Steinke und Ernst Walter Knapp

Veröffentlicht: BMC Bioinformatics 2006, 7:510

Anteil:

- Entwicklung des Konzepts.
- Entwicklung der Operatoren des genetischen Algorithmus.
- Entwicklung des mehrdimensionalen Overlaps .
- Evaluation des Programms (unter anderem mit zirkulär permutierten Testsets. Nicht im Artikel dargestellt, siehe auch Diskussionsteil dieser Arbeit).

Online verfügbar unter:

http://www.biomedcentral.com/1471-2105/7/510

doi:10.1186/1471-2105-7-510

Inhalt des Artikels:

In diesem Artikel wird GANGSTA,eine neue Methode zum Strukturalignment vorgestellt. GANGSTA steht für (**G**enetic **A**lgorithm for **N**onsequential **G**apped **ST**ructure **A**lignment, auf Deutsch: Genetischer Algorithmus für nicht sequenzielles, Lücken zulassendes Strukturalignment)

GANGSTA arbeitet in zwei Stufen, die erste Stufe fokussiert auf Sekundärstrukturelemente (SSE) und ermittelt die strukturelle Ähnlichkeit zwischen zwei Proteinen durch Maximierung einer Fitnessfunktion basierend auf der Zahl möglicher gemeinsamer Paarkontakte zwischen den Strukturelementen und der relativen Orientierung der Strukturelemente. Dazu werden die SSE zunächst entweder durch DSSP[12] oder durch Stride[13] definiert und von C- zu N-Terminus durchnummeriert. Der für GANGSTA entwickelte Genetische Algorithmus stellt die SSE des kleineren Proteins (Quelle) durch ihre Position im Chromosom des Genetischen Algorithmus dar und die zugeordneten SSE des größeren Proteins (Ziel) als die Allelwerte im Chromosom. Dadurch wird das Proteinalignment auf einen Vektor ganzer Zahlen reduziert. Für eine größere Flexibilität werden auch Lücken im Quellprotein zugelassen, aber in der Fitnessfunktion bestraft um möglichst globale Alignments zu gewährleisten. Lücken im größeren Ziel entstehen durch diese Prozedur automatisch und werden nicht bestraft.

In der zweiten Stufe werden die besten Ergebnisse der ersten Stufe auf der Ebene der einzelnen Aminosäureresiduen optimiert. Dabei werden globale und lokale Kriterien gegeneinander aufgewogen. Der GANGSTA Score wird dabei vom RMSD der C_α–Atome, der mit Hilfe des Kabsch Algorithmus[14] bestimmt wird, sowie dem Overlap dominiert. Unter dem Overlap versteht man die Summe der gemeinsamen Residuenpaarkontakte, in Relation zur größeren der Summen der Paarkontakte der beiden einzelnen Proteine. Dabei ist ein Kontakt in GANGSTA folgendermaßen definiert. Zwei Residuen stehen in Kontakt, wenn die Distanz ihrer C_α-Atome kleiner oder gleich 11 Ångström ist.

Zur Evaluation der Güte des beschriebenen Strukturalignmentprogramms werden erstens die Resultate für zwei bekannte Sätze nicht sequenziell alignierbarer Proteine[15] gezeigt. Zweitens wird ein Selbstkonsistenztest zur Überprüfung des Robustheit des Algorithmus beschrieben, in dem gezeigt wird das GANGSTA den Rossman-Fold[16] zuverlässig mit verschiedenen Quellproteinen aus der Datenbank in den selben Zielproteinen wiederfindet. Und drittens werden aus der Literatur die Datensets von Fischer[17] und Novotny[18] herangezogen und die Ergebnisse des Alignments mit GANGSTA mit den

Benchmarkresultaten der von Novotny getesteten Programme (Unter anderem CE[19], DALI[20],SSM[21], VAST[22] und YAKUSA[23]) verglichen.

Circular permuted proteins in the universe of protein folds

Autoren: Tobias Schmidt-Gönner, Aysam Gürler, Björn Kolbeck, Ernst Walter Knapp.

Veröffentlicht: Proteins: Structure, Function, and Bioinformatics 2009 ,78(7):1618

Anteil:

- Entwicklung des Konzepts.
- Mitentwicklung des ursprünglichen Programms zum Strukturalignment.
- Datenbankanalyse und Auswertung der Ergebnisse.
- Manuskript schreiben.

Online verfügbar unter

http://www3.interscience.wiley.com/cgi-bin/fulltext/123218318/HTMLSTART

doi:10.1002/prot.22678

Inhalt des Artikels:

In dieser Arbeit wird eine Analyse der Häufigkeit zirkulär permutierter Proteine im uns bekannten Proteinuniversum vorgenommen. Dazu wurde ASTRAL40[5] (Version 1.73) eine Untermenge des SCOP[24] Datenbank benutzt. In ASTRAL40 befinden sich 9536 Proteindomänen, deren Sequenzidentität unter 40 % liegt.[5,24] Die Datenbasis wurde noch um 351 Proteindomänen verkleinert für die mit GANGSTA+ Symmetrien gefunden wurden, um fehlerhafte zirkuläre Zuordnungen zu vermeiden.[6,7,25] Die übrigen Strukturen wurden dann mit GANGSTA+ auf ihre Strukturähnlichkeit unter besonderer Berücksichtigung zirkulär permutierter Formen untersucht. GANGSTA+ erkennt eine zirkuläre Permutation anhand der Zuordnung der aufeinander abgebildeten SSE (**S**ekundär**s**trukture**l**emente). Für ein sequenzielles Alignment ergibt sich für die Zuordnungen der SSE des Zielproteins zu denen des Quellproteins eine stetig steigende Folge. Im Falle einer Zirkulärpermutation ist diese stetig steigende Folge an exakt einer Stelle unterbrochen. Außerdem muss für eine Zirkulärpermutation die Nummer des letzten zugeordneten SSEs kleiner sein als die des Ersten.

GANGSTA+ erlaubt nicht sequenzielles Strukturalignment in einem ähnlichen Zeitrahmen wie die schnellsten, gängigen sequenziell arbeitende Alignmentprogramme. Auf einem AMD/Opteron mit 1600MHz kann ein Paaralignment in weniger als einer Sekunde berechnet werden. Durch die hohe Geschwindigkeit von GANGSTA+, ist es möglich einen Datenbankscan jeder Proteindomäne gegen alle übrigen durchzuführen. Eine Analyse des jeweils besten Alignments zeigt das zirkuläre Permutationen im Proteinuniversum ziemlich häufig sind. Tatsächlich finden sich für 64 % aller Proteindomänen strukturähnliche Partnerdomänen mit einer zirkulären Permutation. Dies ist ein Ergebnis das zum Teil erheblich höher liegt als die Schätzungen vergleichbarer Arbeiten.[26-28]

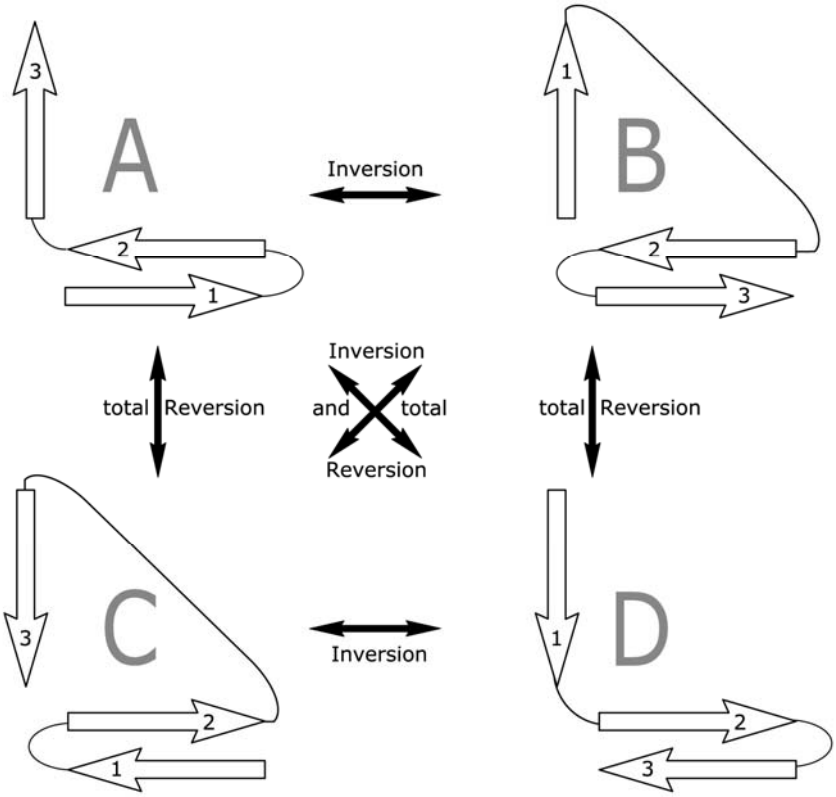


Abbildung 8 Schema der neu eingeführten Alignmentmethoden: Inversion : A->B, totale Reversion:A->C und die Kombination von Inversion und totaler Reversion:A->D

Deshalb werden verschiedene Methoden angewandt um eine Fehlerabschätzung des Ergebnisses zu bekommen. Unter anderem führen wir hierfür verschiedene andere Abbildungsformen für das Strukturalignment ein. Eine Inversion der SSE-Reihenfolge, eine Reversion aller SSE-Leserichtungen, sowie die Kombination der beiden. (siehe Abbildung 8)

Für Proteinpaare, die über diese Strukturalignmentmethoden gefunden werden gibt es keinen in der Natur vorkommenden divergent evolutionären Mechanismus. Dies erlaubt eine Abschätzung des Fehlers durch zufällige Strukturähnlichkeiten, die auf konvergenter Evolution oder gänzlich zufällige Ähnlichkeit zurückgehen, da diese Methoden auf SSE Anordnungen basieren, die laut ihrer Kombinatorik alle die gleiche statistische Wahrscheinlichkeit haben wie das normale Alignment. Die Probleme, die bei dieser vereinfachten Betrachtung entstehen werden im Artikel ausführlich diskutiert.

Zusätzlich wird die in der SCOP Datenbank mitgelieferte Zuordnung zu den Proteinfamilien zur Beurteilung herangezogen, da eine evolutionäre Verwandtschaft außerhalb der eigenen SCOP-Proteinsuperfamilie unwahrscheinlich ist.

Schließlich betrachten wir noch die CPDB, eine erste Datenbank zirkulär permutierter Proteine, die im Laufe dieser Arbeit veröffentlicht wurde. Unglücklicherweise enthält diese Datenbank keine Negativbeispiele, so dass eine bessere Fehlerabschätzung unserer Methode unter Berücksichtigung falsch positiver Bewertungen auch durch Nutzung dieser Datenbank nicht möglich ist, aber sie erlaubt zumindest einen Vergleich zu dem direkt für das Auffinden von zirkulären Permutationen entwickelte Suchwerkzeug CPSARST, das die Grundlage zum Aufbau der CPDB bildet.

Diskussion

Im Rahmen dieser Dissertation wurde ein neues Strukturalignmentprogramm für Proteine vorgestellt. Es ermöglicht im Gegensatz zu gängigen bekannten Alignmentprogrammen wie Dali[20,29] oder K2SA[30] auch ein Proteinstrukturalignment außerhalb der sequenziellen Verknüpfung der Residuen. Zudem wurden zwei Anwendungsgebiete vorgestellt, für die es notwendig war eben diese sequenzielle Verknüpfung zu verlassen. Zum einen wurde der Rossmann-Fold[16] untersucht, ein Strukturmotiv das Proteinen als Bindungsstelle für ATP und verwandte Energieträger der Zelle dient. Es wurde dabei gezeigt, dass dieses Strukturmotiv in verschiedenen, nicht sequenziellen Anordnungen existiert. GANGSTA ermöglicht das Auffinden solcher Motive.
Zum zweiten wurden zirkulär permutierte Proteine untersucht, eine Verwandtschaft die einen Bruch in der Sequenz voraussetzt.
Aus der Literatur als zirkulär permutiert bekannte Proteinpaare bildeten eines der ersten Testsets, die für GANGSTA zusammengestellt wurden. An ihnen lässt sich die Leistung von GANGSTA gut mit denen anderer Strukturalignmentprogramme vergleichen, da auch gängige sequenzielle Alignmentprogramme diese Proteinpaare bearbeiten können, auch wenn diese immer nur einen Teil des Proteins zuordnen können. Gleichzeitig eignen sie sich zum Test der nicht sequenziellen Lösungen unserer Methode.

Tabelle 1: Testset zirkulär permutierter Proteine.

ZPP	Name (PDB code)	SCOP-Klasse	Herkunft
1	Pea lectin (1rin)	All Beta	Erbse
1	Concanavalin A (2cna)	Mainly Beta	Jackbohne
2	Nk-lysin (1nkl)	All Alpha (NMR)	Schwein
2	Prophytepsin (1qdm)	All Alpha	Gerste
3	Synaptotagmin C2-Domäne (1rsy)	All Beta	Ratte
3	Phosphodiesterase (1qas)	All Beta	Ratte
4	DNA methyltransferase M.Taq I(1aqi)	Alpha and Beta	Thermus aquaticus
4	pvu II DNA methyltransferase(1boo)	Alpha and Beta	Proteus vulgaris
5	transaldolase B(1onr)	Alpha and Beta	Escherichia coli
5	fructose-1,6-bisphosphate aldolase(1fba)	Alpha and Beta	Fruchtfliege
6	beta-glucanase(1gbg)	All Beta	Bacillus licheniformis
6	CPA16M-84(1ajk) [Künstlich]	All Beta	Paenibacillus macerans
7	avidin(1avd)	All Beta	Huhn
7	E51/A46(1swg) [Künstlich]	All Beta	Streptomyces avidinii

Das Testset (siehe Tabelle 1) enthält Beispiele verschiedener SCOP-Klassen, und sowohl natürlich vorkommende Vertreter wie speziell künstlich hergestellte Zirkulärpermutationen.

Dabei zeigt sich, dass GANGSTA in der Lage für alle getesteten Paare die Zirkulärpermutation zu erkennen, aber für die Paare 1aqi-1boo und 1onr-1fba sind die Alignments von einer zu geringen Qualität um eine sinnvolle Beurteilung zu erlauben. Für 1aqi-1boo ist die Zahl der Lücken größer als die Zahl der zugeordneten Sekundärstrukturelemente und für 1onr-fba beträgt der RMSD 4 Ångström. Die Zuordnungen und die Qualität kann den Tabellen 2 und 3 entnommen werden.

Tabelle 2: GANGSTA-Alignments bekannter Zirkulär permutierter Proteinpaare.

ZPP	GANGSTA-Alignment	ZP gefunden
1rin – 2cna	9,10,-,-,12,13,1,2,3,4,5	+
1nkl – 1qdm	3,4,1,2	+
1rsy – 1qas	8,1,2,-,4,-,5,6,-	+
1aqi – 1boo	-,-,6,-,9,-,11,12,-,-,-,2,-,4,-,-	±
1onr – 1fba	-,19,-,20,1,2,3,4,-,-,-,-,7,8,9,11,14,16,17,18,-	±
1gbg – 1ajk	11,12,13,14,15,16,2,3,4,5,6,7,8,9,10	+
1avd – 1swg	7,8,9,1,2,3,4,6	+

Tabelle 3: Vergleich verschiedener Alignmentprogramme.

	DaliLite[29]		K2SA[30]		MAMMOTH[31]		GANGSTA[6]	
	# Res	RMSD	# Res	RMSD	# Res	RMSD	# Res	RMSD
1rin – 2cna	106	1.7	107	1.0	24	3.8	123	0.9
1nkl – 1qdm	55	2.7	29	2.6	27	3.7	55	2.8
1rsy – 1qas	109	3.7	89	1.1	74	2.9	82	1.0
1aqi – 1boo	113	3.9	57	2.2	22	3.8	104	2.9
1onr – 1fba	198	4.1	81	2.3	54	3.8	147	4.0
1gbg – 1ajk	123	1.2	116	0.4	119	1.4	194	0.5
1avd – 1swg	74	1.7	68	1.0	57	2.8	85	0.7

#Res = Zahl der zugeordneten Residuen, RMSD = RMSD der C_α-Atome in Å.
Grau unterlegt sind die Felder mit dem jeweils besten Ergebnis, getrennt nach #Res und RMSD, da die Wichtung der beiden Qualitätsfaktoren je nach Programm schwanken kann.

Die Alignmentprogramme DaliLite[29] und K2SA[30] lieferten von den 2007 getesteten Programmen neben GANGSTA die besten Ergebnisse bei dieser Aufgabe, aus der Riege anderer getesteter Strukturalignmentprogramme wurde für diese Übersicht noch

MAMMOTH[31] ausgewählt, weil es wie GANGSTA zusätzliche Funktionen erfüllen soll, und deshalb wie GANGSTA einen größeren Suchraum bearbeitet. Bei MAMMOTH liegt der Fokus jedoch auf Multiplen Strukturalignments. Trotz der Tatsache, dass die sequenziell arbeitenden Programme nur einen Teil des Proteins alignieren konnten, liegen sie in einem ähnlichen Bereich wie GANGSTA. Das liegt daran, dass GANGSTA die Random Coil Bereiche ignoriert, diese bei ähnlichen Proteinpaaren aber ebenfalls eine hohe Ähnlichkeit aufweisen können und deshalb zu den Alignments der anderen Programme beitragen können.
Mit GANGSTA lag also schon 2007 ein Programm vor, das im paarweisen Alignment zirkulär permutierte Proteine ziemlich zuverlässig erkennen konnte und im Vergleich mit anderen Programmen gleichwertige Ergebnisse liefert.
Für die groß angelegten Datenbankscans, wie sie für die Untersuchung zirkulär permutierter Proteinpaare notwendig waren, hatte GANGSTA noch keine ausreichende Performanz. Dieses Problem konnte erst mit der Einführung von GANGSTA+ gelöst werden. GANGSTA+ ist mehr als zehn mal schneller als GANGSTA. Dabei wurde die hierarchische Strategie des ursprünglichen Lösungsansatzes beibehalten, aber die einzelnen Stufen wurden methodisch überarbeitet. Anstelle des genetischen Algorithmus werden die Alignments auf der SSE Ebene nun durch eine iterative Kombination von SSE-Tupeln erzeugt, ein Ansatz aus der dynamischen Programmierung. Die dabei entstehenden Strings entsprechen den Chromosomen des genetischen Algorithmus. Auf der zweiten Stufe wird jetzt eine Energieminimierung durchgeführt, bei der zwischen C_α-Atomen die zu unterschiedlichen Proteinen gehören eine leicht anziehende Kraft benutzt wird. Die beiden Proteine werden dabei als starre Körper behandelt. Am Ende dieser Energieminimierung werden die C_α-Atome, die auf einem Gitter denselben Punkten zugeordnet werden, aligniert. Dadurch können bei GANGSTA+ im Nachhinein auch Residuen außerhalb der anfänglich bestimmten SSEs einander zugeordnet werden.

Bisher sind zirkulär permutierten Proteinpaare nur durch speziell für diese Aufgabe programmierte Anwendungen auffindbar gewesen. Mit GANGSTA+ liegt nun ein universelles Strukturalignmentprogramm vor, das eine solche Untersuchung ebenfalls erlaubt. Zusätzlich ermöglicht GANGSTA+ verschiedene Strukturverwandtschaften zu definieren, die nicht auf natürlichen Verwandtschaften zurückzuführen sind. Diese künstlichen Strukturverwandtschaften eröffnen einen neuen Weg den Fehler bei der datenbankbasierten Suche nach zirkulär permutierten Proteinen abzuschätzen.
Wichtigstes Ergebnis ist die Einschätzung, dass deutlich mehr zirkulär permutierte Proteine existieren als bisher angenommen.

Zusammenfassung

Im Rahmen dieser Arbeit wurde das nicht sequenziell arbeitende Proteinstrukturalignmentprogramm GANGSTA vorgestellt und auch sein Nachfolger GANGSTA+ diskutiert. Es wurde gezeigt das beide Programme, trotz des größeren Suchraums, der für nicht sequenzielle Strukturalignments abgetastet werden muss, eine mit etablierten sequenziell arbeitenden Strukturalignmentprogrammen vergleichbare Leistung bei Alignmentqualität und Klassifizierungsfähigkeit von Proteinen erzielen. Durch den Geschwindigkeitsgewinn gegenüber dem Vorgänger ist GANGSTA+ auch bei der rechnerischen Performance zu sequentiell arbeitenden Strukturalignmentprogrammen gleichwertig. Es wurde außerdem nachgewiesen, dass durch die Berücksichtigung nicht sequenzieller Strukturalignments auch Probleme in der Biochemie behandelt werden können, die sequentiell nicht lösbar sind. Dies ist exemplarisch anhand der vielfältig sequentiell unterschiedlichen Strukturalignments des Rossmann-Folds für ein biologisch relevantes Proteinstrukturmotiv gezeigt worden. Mit der Analyse der Häufigkeit zirkulär permutierter Proteine im uns bekannten Proteinuniversum ist ein wichtiger Typ evolutionärer Verwandtschaften untersucht worden, die sich konventionellen Methoden des Strukturalignments verschließt. Die gefundene Zahl der zyklisch permutierten Proteinpaare erweist sich dabei als höher als bisher angenommen.

GANGSTA+ ist unter http://agknapp.chemie.fu-berlin.de/gplus/ für die Öffentlichkeit verfügbar.

English Summary

In this work, GANGSTA, a non-sequentially working protein structure alignment tool, was introduced and his successor GANGSTA+ was shortly discussed. It has been shown that both programs are comparable to established sequentially working structure alignment tools, in quality and in their ability to classify the proteins in different super families, although non-sequential alignments need to explore a much larger search space. Due to the performance enhancement of GANGSTA+ calculation times are in the same range as sequentially working tools.

It was shown that including non sequential alignments allows GANGSTA to tackle certain biological problems, which can not be solved by using programs that are restricted to sequential alignments only. This has been exemplified by the non sequential alignments of the Rossmann folds, a biological relevant structural motif involved in energy transfer in the cell.

The analysis of the occurrence of circular permuted proteins in the known universe of protein structure detected a larger number of circular permuted protein pairs than expected.

GANGSTA+ is available at http://agknapp.chemie.fu-berlin.de/gplus/.

Literaturverzeichnis

1. Muirhead H, Perutz M. Structure of hemoglobin. A three-dimensional fourier synthesis of reduced human hemoglobin at 5.5 Å resolution. Nature 1963;199:633-638.
2. Kendrew J, Bodo G, Dintzis H, Parrish R, Wyckoff H, Phillips D. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature 1958;181:662-666.
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235-242.
4. Anfinsen C. Principles that govern the folding of protein chains. Science 1973;181:223-230.
5. Chandonia JM, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE. The ASTRAL compendium in 2004. Nucleic Acids Res 2004;32:D189 - D192.
6. Kolbeck B, May P, Schmidt-Goenner T, Steinke T, Knapp EW. Connectivity independent protein-structure alignment: a hierarchical approach. BMC Bioinformatics 2006;7(1):510.
7. Guerler A, Knapp EW. Novel protein folds and their nonsequential structural analogs. Protein Sci 2008;17(8):1374-1382.
8. Wallace CJA. The curious case of protein splicing: Mechanistic insights suggested by protein semisynthesis. Protein Sci 1993;2:697-705.
9. Heinemann U, Hahn M. Circular permutations of protein sequence: not so rare? Trends Biochem Sci 1995;20(9):349-350.
10. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48(3):443 - 453.
11. Smith TF, Waterman MS. Identification of Common Molecular Subsequences. J Mol Biol 1981;147:195-197.
12. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577-2637.
13. Frishman D, Argos P. Knowledge-based secondary structure assignment. Proteins 1995;23:566-579.
14. Kabsch W. A solution for the best rotation to relate two sets of vectors. Acta Cryst 1976;A32:922-923.
15. Dror O, Benyamini H, Nussinov R, Wolfson HJ. Multiple structural alignment by secondary structures: Algorithm and applications. Protein Sci 2003;12(11):2492-2507.
16. Rossmann MG, Moras D, Olsen KW. Chemical and biological evolution of a nucleotide-binding protein. Nature 1974;250:194-199.
17. Fischer D, Elofsson A, Rice D, Eisenberg D. Assessing the performance of fold recognition methods by means of comprehensive benchmark. Proc Pacific Symposium on Biocomputing 1996:300-318.
18. Novotny M, Madsen D, Kleywegt GJ. Evaluation of Protein Fold Comparison Servers. Proteins 2004;54:260-270.
19. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 1998;11:739 - 747.
20. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123-138.
21. Krissinel E, Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 2004;60(Pt 12 Pt 1):2256 - 2268.
22. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6(3):377-385.

23. Carpentier M, Brouillet S, Pothier J. YAKUSA: A fast structural database scanning method. Proteins 2005;61:137-151.
24. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536-540.
25. Guerler A, Wang C, Knapp E-W. Symmetric Structures in the Universe of Protein Folds. J Chem Inf Model 2009;49(9):2147-2151.
26. Lo WC, Lyu PC. CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships. Genome Biol 2008;9(1):16.
27. Uliel S, Fliess A, Amir A, Unger R. A simple algorithm for detecting circular permutations in proteins. Bioinformatics 1999;15(11):930-936.
28. Abyzov A, Ilyin VA. A comprehensive analysis of non-sequential alignments between all protein structures. BMC Struct Biol 2007;7(1):78.
29. Holm L, Park J. DaliLite workbench for protein structure comparison. Bioinformatics 2000;16:566-567.
30. Szustakowski JD, Weng Z. Protein structure alignment using a genetic algorithm. Proteins 2000;38:428-440.
31. Ortiz AR, Strauss CEM, Olmea O. MAMMOTH (Matching molecular models obtained from theory): An automated method for model comparison. Protein Sci 2002;11:2606-2621.