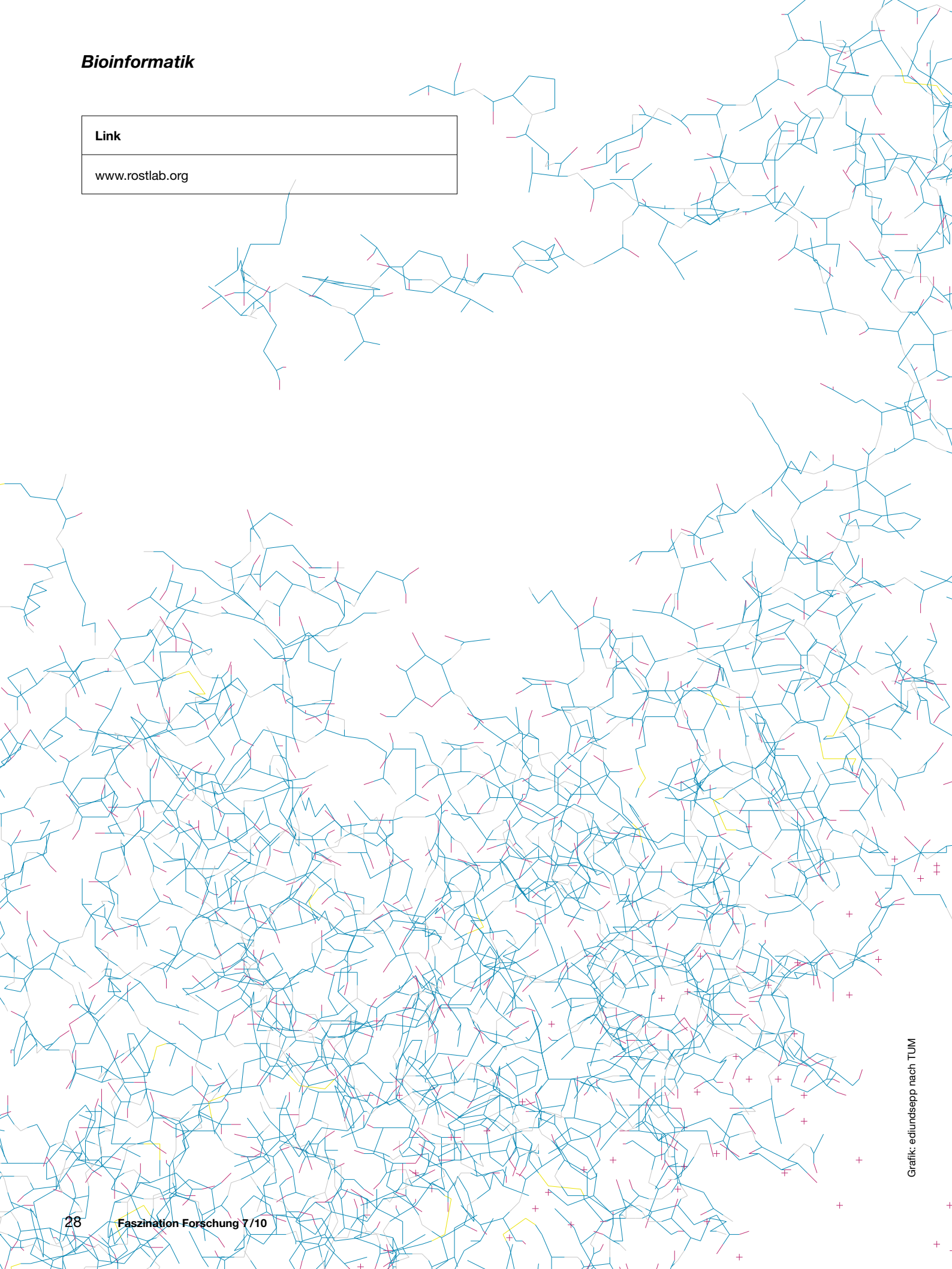
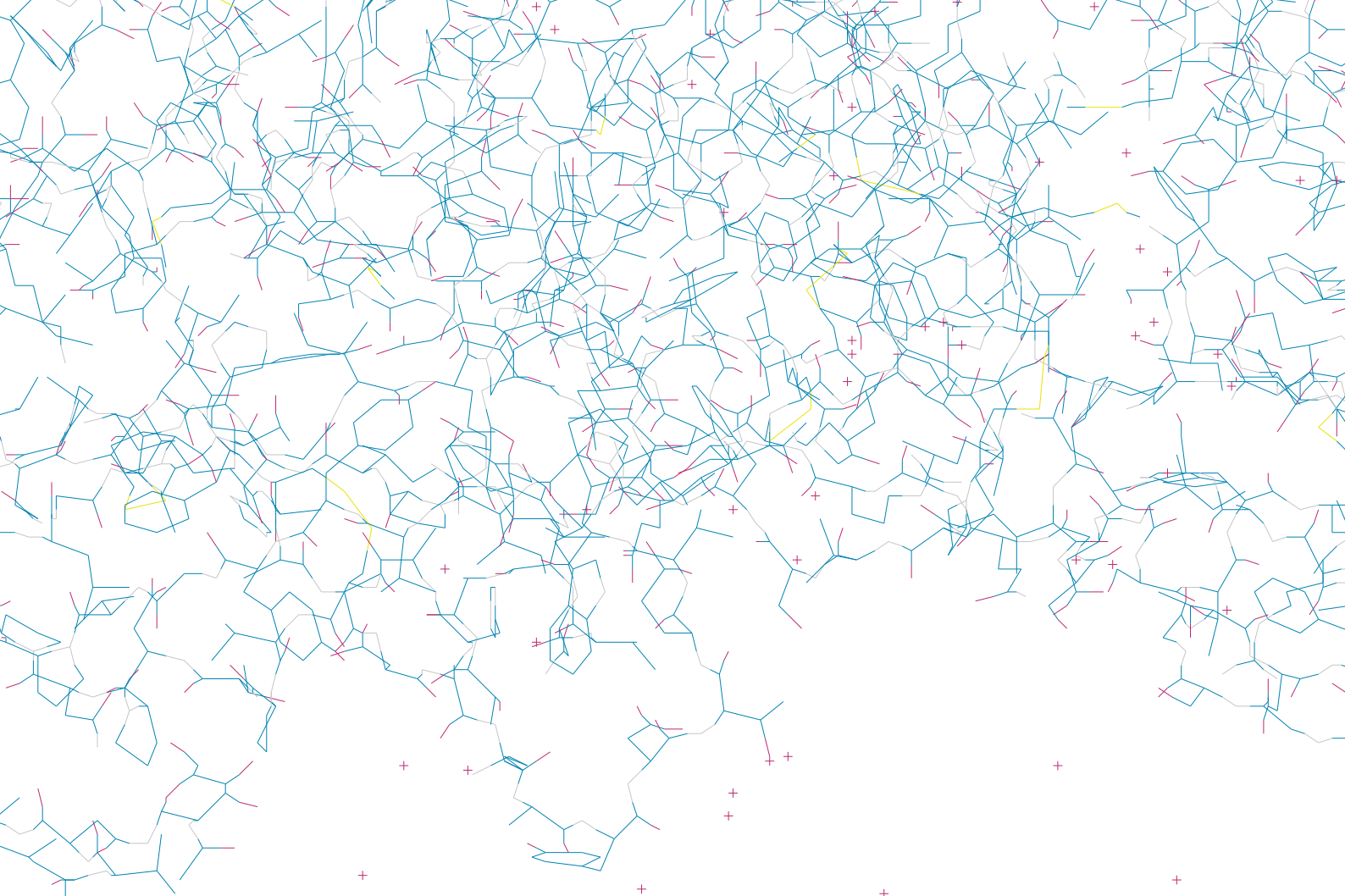
Bioinformatik

Link
www.rostlab.org



Grafik: ediundsepp nach TUM



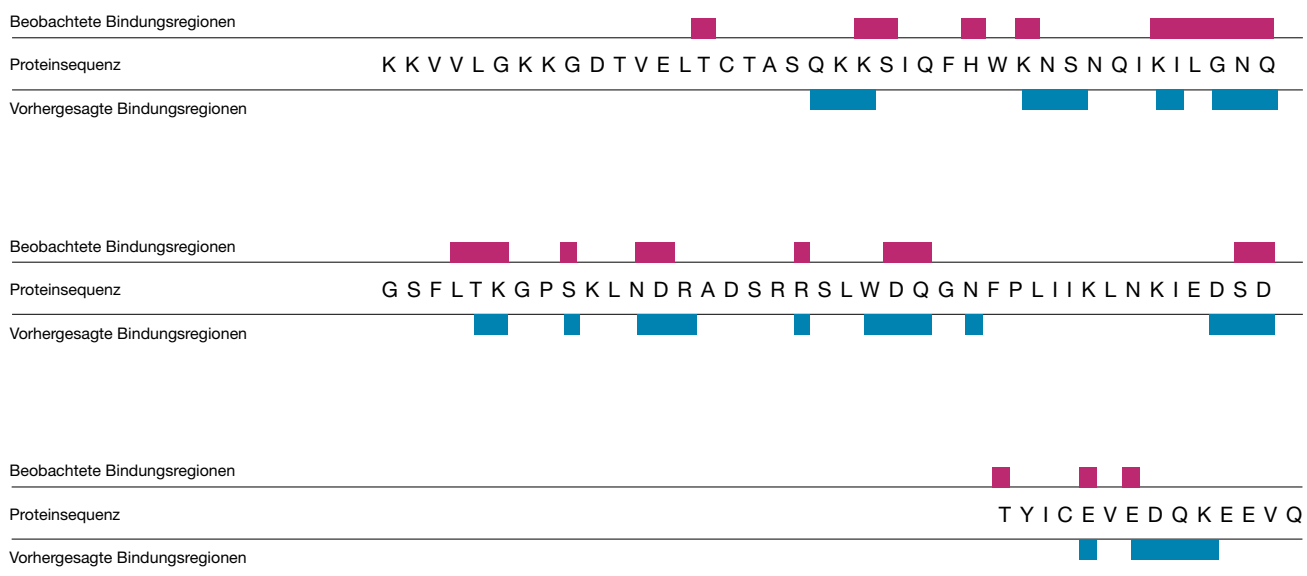
Leben berechnen

Bioinformatiker treffen mit Methoden des maschinellen Lernens und neuronalen Netzen Vorhersagen über die Strukturen von Proteinen und deren Funktion

Die Erwartungen waren hoch, als im Mai 2003 offiziell verkündet wurde, das menschliche Genom sei vollständig entschlüsselt. Endlich, so die Hoffnung, könne man Krankheitsursachen aufklären, die auf Gendefekten beruhen, und Therapien entwickeln. „Was bedeutet entschlüsselt? Wir kennen die genaue Abfolge der menschlichen DNA mit sechs Milliarden Bausteinen, also den Basen Adenin, Thymin, Cytosin und Guanin. Diese Erbgutkarte lässt sich mit einem Benutzerhandbuch vergleichen. Ich habe die Erfahrung gemacht, dass ich in solchen Handbüchern selten das finde, was ich suche. Und wenn ich es finde, verstehe ich es nicht“, schildert Prof. Burkhard Rost vom TUM-Lehrstuhl für Bioinformatik, dessen Arbeitsgruppe sich mit der Speicherung, Analyse und Interpretation von Gen- und Proteindaten beschäftigt, das Problem.

Von Therapien ist die Wissenschaft immer noch weit entfernt. Was die Forscher wissen: Das menschliche ▷

Vorhersage von Bindungsregionen an Proteinen



Grafik: ediundsepp nach TUM

Genom besteht aus 25.000 Genen und nicht wie ursprünglich gedacht aus 100.000. Diese Gene enthalten die Bauanweisung für die Proteine. Zum Vergleich: Die Reispflanze besitzt 50.000, ein Wurm 21.000 und Mycoplasma genitalium, der kleinste bekannte Organismus, 470 Gene.

Regler für alle Lebensprozesse

Ohne Proteine kann ein biologisches System nicht existieren. „Maschinerie des Lebens" nennt Rost sie. Denn Proteine transportieren Substanzen, sorgen für die Blutgerinnung, agieren als Ionen-Pumpen, beschleunigen chemische Reaktionen oder erkennen Signalstoffe. Kurz: Sie regeln alle Lebensprozesse. Fehler in ihrem Aufbau oder in ihrer Funktion machen den Menschen krank.

In ihrer Ausgangsform liegen Proteine als kurze oder lange Aminosäure-Ketten vor (Primärstruktur). Abhängig von ihrer chemischen Zusammensetzung ordnen sich die Aminosäuren zu zweidimensionalen Gebilden (Sekundärstruktur) an. Je nach Kräfteverhältnissen und Bindungen falten sie sich weiter (Tertiärstruktur) – dann sehen sie aus wie ein Wollknäuel. Manche Proteine müssen sich sogar zu Komplexen zusammenlagern (Quartiärstruktur), um funktionieren zu können. So bestehen beispielsweise Antikörper (Immunglobuline) aus vier Proteinen. Ihre Funktion wird durch die dreidimensionale Anordnung ihrer Atome bestimmt.

Die Natur macht es den Forschern nicht leicht: „Wir wissen immer noch nicht genau, wie viele Proteine der Mensch hat. Wir glauben, es sind rund 25.000. Von mehr als der Hälfte ist die Funktion immer noch unbekannt. Das heißt, wir verstehen das Grundprinzip des Lebens noch nicht", stellt der Bioinformatiker fest. Aber das ist noch nicht alles: Über 80 Prozent der menschlichen Proteine haben mehr als eine Funktion. Und mittlerweile ist auch bekannt, dass die meisten Proteine als Paare oder größere Komplexe auftreten.

Die Faltung der dreidimensionalen Strukturen hängt von vielen Faktoren ab: etwa vom pH-Wert oder vom Bindungspartner, ebenso kann das Binden selbst Funktion und Struktur verändern. „Proteine sind nicht statisch. Das macht das Ganze so kompliziert. Man kann zwar mit heutigen Techniken Proteine identifizieren und die atomare Struktur darstellen, nicht aber ihr dynamisches Verhalten", erklärt Burkhard Rost.

Beobachtete Bindungsregionen

Vorhergesagte Bindungsregionen

Die Forscher nutzen computergestützte Methoden, um die Bindungsregionen (Hot Spots) eines Proteins mit einem Bindungspartner vorherzusagen. Dazu vergleichen sie Proteinsequenzen und suchen darin die Bindungsregionen. Das Beispiel veranschaulicht die vorhergesagten (blau) Bindungsregionen des Proteins CD4, das auf verschiedenen Zellen des menschlichen Immunsystems zu finden ist und vergleicht diese mit den Bindungsregionen (magenta), die in einem Komplex mit dem Protein GP120, das sich auf der Oberfläche von HI-Viren befindet, beobachtet werden. Die Forscher gewinnen durch diese Analysen der Struktur und Funktion der Proteinsequenzen wichtige Hinweise auf die Angriffspunkte, die HI-Viren an den Immunzellen nutzen, wenn sie diese infizieren.

Bilder: TUM

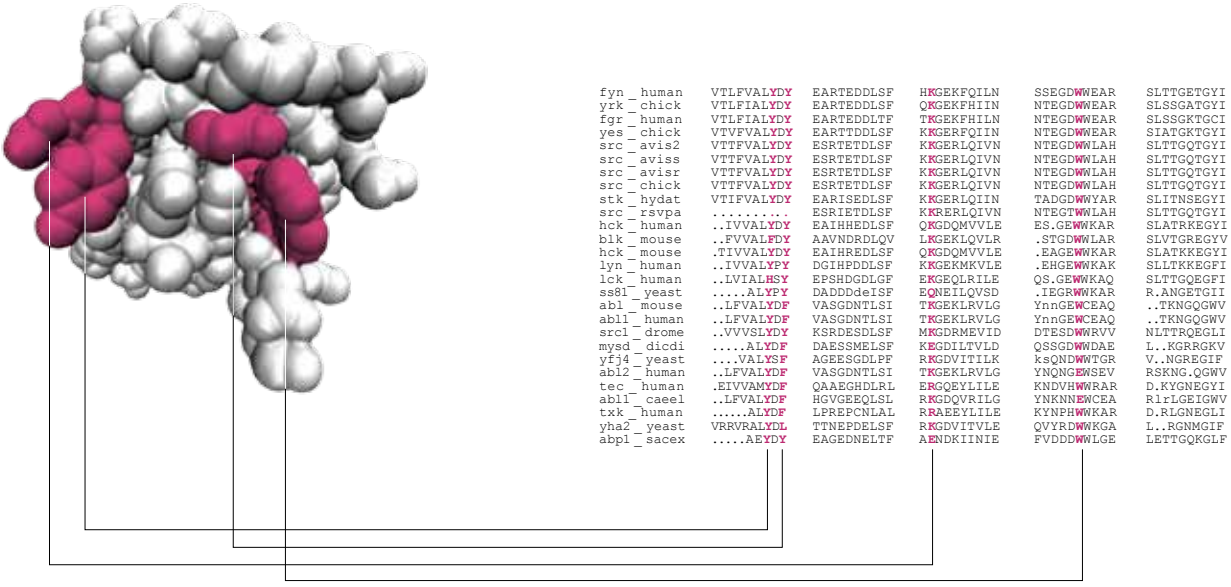
Struktur und Funktion aufklären

In den vergangenen Jahren haben zahlreiche Forscher die Strukturen Tausender Proteine und Proteinkomplexe aufgeklärt und in Datenbanken gespeichert. Sie haben Rechenprogramme entwickelt, um die gesammelten Gen- oder Proteindaten zu vergleichen, und Programme, die Proteine mit ähnlicher Sequenz und biologischer Funktion zu „Familien“ zusammenstellen. Und sie haben herausgefunden, dass etwa ein Drittel aller derzeit in Datenbanken gespeicherten Proteinsequenzen Ähnlichkeiten mit der Sequenz von mindestens einer aufgeklärten Proteinstruktur aufweisen.

Um die Aufklärung zu beschleunigen, versuchen die Bioinformatiker mit computergestützten Methoden Bindungsregionen und physikalische Wechselwirkungen vorherzusagen. „Wir nutzen dazu die Informationen, die uns die Evolution gibt. Das heißt, ähnliche DNA-Sequenzen geben Hinweise auf Verwandtschaft und Abstammung von Organismen. Ist die Sequenz identisch, ist auch die Proteinstruktur gleich. Wenn wir wissen, welche Änderungen möglich sind, dann können wir auch Aussagen über die Struktur und Funktion treffen“, erklärt Burkhard Rost.

Maus und Mensch haben mehr als 97 Prozent ihrer Gene gemeinsam. Diesen Umstand nutzen die Forscher, um aus Übereinstimmungen mit den Maus-Daten auf Struktur und Funktion der menschlichen Proteine zu schließen. Würden die Forscher versuchen, Funktion und mögliche Bindungspartner experimentell herauszufinden, wären sie Jahrzehnte beschäftigt. Deshalb ist es notwendig, ein automatisiertes Verfahren zu entwickeln. „Können wir am Rechner vorhersagen, wie Protein A mit Protein B reagiert? Können wir vorhersagen, welche Regionen sogenannte Hot Spots sind, wie sie wechselwirken?“, lauten die Fragen, die Rost und seine Mitarbeiter beschäftigen.

Dazu nutzt der Lehrstuhl für Bioinformatik Methoden des maschinellen Lernens. Vereinfacht gesagt, „lernt“ der Computer aus vorgegebenen Beispielen, Muster zu erkennen. Dabei soll die Maschine automatisch bekannte Muster in neuen Datensätzen wiederfinden. Die Umsetzung geschieht mittels entsprechender Rechenanweisungen. Solche Algorithmen werden bereits bei Briefsortieranlagen eingesetzt, wenn es darum geht, automatisch Postleitzahlen auf Briefen zu erkennen. Mit dieser Methode sind auch Ärzte bei der ▷



Grafik: ediundsepp nach TUM

Aminosäuren, die im Molekül nahe beieinanderliegen (magenta markierte Regionen), sind gekoppelt: Verändert sich die eine Aminosäure, hat dies auch Auswirkungen auf die andere. So können die Forscher aus dem Vorhandensein einer bestimmten Aminosäure gewisse Rückschlüsse auf die Entwicklung des Proteins ziehen

Diagnose zu unterstützen oder Fahrzeugleitsysteme so zu entwerfen, dass sie die Strecke unter Berücksichtigung der zu erwartenden Auslastung optimieren.

Neuronale Netze als Vorbilder

Rosts Mitarbeiter setzen ebenso „Trainingsmethoden" nach dem Vorbild neuronaler Netze ein: Der Rechner soll durch Mustererkennung Bindungsbereiche auf den Proteinoberflächen vorhersagen. Nach Abschluss der Trainingsphase sollte es dem neuronalen Netzwerk möglich sein, die Bindungsstellen in zuvor nie gesehenen Proteinen zu identifizieren und zu klassifizieren.
Rosts Labor geht noch einen Schritt weiter: Für die Veränderung von Proteinstruktur und -funktion genügt bereits eine Punktmutation, also der Austausch einer Base, auf der DNA. Seine Mitarbeiter entwickeln derzeit eine Methode, um die Effekte einer solchen Nukleotid-Veränderung (Single Nulear Polymorphism, SNP) vorherzusagen. „Wir versuchen dadurch, mehr über die Mobilität der Proteine zu erfahren. Und wir wollen auch wissen, welche Aminosäure-Reste auf welche Weise verändert werden können. Heute sind wir nicht in der Lage, das unter dem Mikroskop zu sehen. Aber wir könnten es simulieren und versuchen, unsere Vorhersagen zu bestätigen", erläutert Rost.
Es gäbe noch einen weiteren Vorteil: Wenn die Wissenschaftler herausfinden, wie die Proteine reagieren, ließen sich Medikamente entwickeln, welche an den Hot Spots wirken. Dies könnte in Zukunft die Behandlung von Krebspatienten erleichtern. Sie würden ein auf ihr individuelles DNA-Profil maßgeschneidertes Medikament erhalten. Denn gerade in der Tumortherapie sprechen die Patienten sehr unterschiedlich auf Wirkstoffe an.

Daten-Explosion in der Protein-Datenbank

Die Rost-Gruppe ist eines von vielen Teams, die weltweit daran arbeiten, computergestützte Vorhersagen zu entwickeln. Alle sind sie auf experimentelle Daten über Proteinstrukturen und -funktionen angewiesen. Die Zahl der experimentellen Daten wächst kontinuierlich. So enthält beispielsweise die frei zugängliche Protein Data Bank (PDB) mehr als 50.000 solche Strukturen. Drei Viertel davon wurden in den letzen fünf Jahren aufgeklärt.
Trotz dieser Steigerung entfallen auf eine ermittelte Proteinstruktur 50 bis 100 unbekannte Strukturen. Unge-

Foto: Eckert/Heddergott, TUM

Burkhard Rost hat die Entstehung der Bioinformatik als Wissenschaftszweig maßgeblich mitgestaltet. 2008 kam er von der Columbia University an die TUM

Geschichte

Die Anfänge der Bioinformatik reichen bis zum „Atlas über Proteinsequenzen und -strukturen“ (erschienen 1965) zurück. Dieser enthielt nicht mehr als zwei Dutzend Proteinsequenzen – darunter das Hormon Insulin – und war die erste Proteindatenbank. Heute existieren weit mehr als 500 biologische Datenbanken. Die drei wichtigsten Banken für DNA-Sequenzen finden sich in Europa, Japan und den USA.

Der Begriff Bioinformatik umfasste anfangs (Mitte der 1980er-Jahre) ursprünglich die Robotertechnik und künstliche Intelligenz. Heute versteht man darunter die Speicherung, Analyse und Interpretation von Gen- und Proteindaten. Das Fach vereinigt die Disziplinen Informatik, Mathematik, Statistik, Physik sowie Methoden aus den Ingenieurwissenschaften mit der Molekularbiologie.

achtet der enormen Daten-Explosion wird also die Kluft zwischen dem, was die Forscher wissen, und dem, was sie wissen möchten, immer größer. Ein typisches Beispiel ist die DNA-Sequenzierung von gesamten Organismen: Über zehn Jahre Arbeit investierte die Wissenschaft in das erste menschliche Genom, heute können große Institute die DNA von zehn Menschen an einem Tag sequenzieren.

Wegbereiter für die individuelle Medizin

Diese Tatsache ist vor allem für das Gebiet der „individuellen“ Medizin relevant. „In fünf bis zehn Jahren liegen unsere Genomprofile dem behandelnden Arzt vor. Wir hoffen, dass wir aus den kleinen Unterschieden zwischen uns, den SNPs, lernen können, welche Medikamente am besten für jedes Individuum geeignet sind. Die SNP-Daten lassen sich schnell ermitteln, die Herausforderung besteht darin, dass Wissenschaft und Medizin mit dieser Datenflut fertigwerden müssen. Eines ist bereits heute klar: Zu Beginn werden Patienten nicht die bestmögliche Behandlung erhalten können, weil uns einfach die notwendigen Computer und Algorithmen im Moment fehlen“, befürchtet Burkhard Rost.

Es sei bezeichnend, meint Rost, dass dies eines der ersten Felder in den Zukunftswissenschaften ist, in dem die typischen Trendsetter in Technik und Wissenschaft wie die USA, Großbritannien und Deutschland in mancher Hinsicht von Asien überholt worden seien. So wurde das Beijing Genomics Institute, ursprünglich als die weltweit größte Sequenzieranlage gebaut, nochmals sehr viel größer dimensioniert, indem es die Verantwortlichen nach Shenzhen (bei Hongkong) verlegten. Rost und seine Kollegen würden gerne entsprechende Anlagen in München etablieren. In solch einem Forschungszentrum könnten Mediziner, Biologen, Bioinformatiker und herausragende Wissenschaftler anderer Zukunftsfelder zusammenarbeiten, „damit Europa in der Forschung nicht zurückfällt“.

Können wir also – wenn die Voraussetzungen stimmen – bald Leben berechnen? Burkhard Rost muss nicht lange überlegen: „Nein, bis jetzt sind wir nicht dazu in der Lage. Zu viel ist noch unbekannt. Aber wir haben einen enormen Wissensstand erreicht, und wir finden jeden Tag mehr. Was wir jetzt schon berechnen können, hilft uns jeden Tag. Wir können hoffen.“

Evdoxía Tsakiridou