

VOM SCHLÜSSEL zur Funktion



Genforschung in der GSF



Bioinformatik als Genom

Organismen bestehen aus Tausenden von Proteinmolekülen, die von Millionen oder auch Milliarden von Basenpaaren der DNA kontrolliert abgelesen werden. Damit ergibt sich ein Informationsgehalt der für das menschliche Auffassungsvermögen kaum begreifbar ist. Mit Hilfe von Informatik, Statistik und Mathematik gehen Wissenschaftler am Institut für Bioinformatik neue Wege auf der Suche nach Lösungen für die Analyse und Interpretation dieser unfassbar großen Datenbestände.

*Du bleibst zu Hause, Wichtigstes zu tun.
Entfalte du die alten Pergamente,
Nach Vorschrift sammle Lebenselemente
Und fuege sie mit Vorsicht eins ans andre.
Das Was bedenke, mehr bedenke Wie.
Indessen ich ein Stueckchen Welt durchwandre,
Entdeck' ich wohl das Tuepfchen auf das i.*

J. W. v. Goethe, Faust. Der Tragödie zweiter Teil

Das **Buch des Lebens** ist in gedruckter Form nicht lesbar, es verschließt sich der Interpretation. Zu allererst stellt sich ein quantitatives Problem: Während der Text des zweiten Teil der Fausttragödie in 300 Kilobyte abgespeichert werden kann, ist das Hefegenom 40 mal, das Humangenom etwa 9000 mal so umfangreich.

Einer der großen Zufälle in der Geschichte der Naturwissenschaften

ist die Parallelität der Entwicklung der Molekularbiologie und der Computertechnologie. Ohne leistungsfähige Rechner ist eine systematische Analyse der Genominformationen nicht möglich. Allein der Vergleich von 1000 Proteinsequenzen untereinander erfordert die Auflösung einer halben Million von Optimierungen und deren statistische Bewertung. Trotzdem es derzeit über eine Million Sequenzen in den Datenbanken gibt, kann die Bioinformatik mit den verfügbaren Rechnern und optimierten Methoden alle diese Vergleiche durchführen. Die Fragen, mit welchen sich Hans Werner Mewes und seine Mitarbeiter am Institut für

Bioinformatik beschäftigen, gehören zwei unterschiedlichen Klassen an: Einerseits geht es um die Anwendung von Algorithmen, Methoden und Werkzeugen der Bioinformatik auf experimentelle Daten zum Einblick in biologische Zusammenhänge. Andererseits nutzen sie Daten zur Entwicklung von Methoden, meist verbunden mit theoretischen Überlegungen, die zur Entwicklung neuer Algorithmen führen. Die Genvorhersage in eukaryontischen Genomen liefert dafür ein geeignetes Beispiel. Liegt die Sequenz eines Genoms vor, besteht die erste und wichtigste Aufgabe in der Identifizierung der genetischen Elemente, also der funktionel-

Werkzeug Genomanalyse

len Abschnitte der DNA-Sequenz. Dazu gehören die kodierenden Regionen, also die proteinogenen Abschnitte ebenso wie die regulatorisch aktiven Sequenzen. Höhere Eukaryonten verfügen über komplexe Mechanismen, die die primär transkribierte RNA-Sequenz prozessieren und dabei bestimmte Abschnitte der RNA, die Introns, ausschneiden. Die Primärstruktur von Proteinen kann nur dann korrekt vorhergesagt werden, wenn es gelingt, diesen Prozess detailgenau zu modellieren.

Struktur und Kreativität - zwei Herzen in der Brust des Bioinformatikers

Während die Biologie nur an der Sequenz des Proteins interessiert ist, bleibt Hans Werner Mewes die Aufgabe, dafür den besten Algorithmus zu entwickeln, also eine möglichst zuverlässige Vorhersage zu erreichen. Damit kommt er selbstkritisch schnell zum Dilemma und zur Herausforderung der Bioinformatik: „Solange der Prozess nicht perfekt modelliert werden kann, ist es erforderlich, die Vorhersage experimentell zu überprüfen oder auch die Modelle unterschiedlicher Algorithmen manuell zu korrigieren und so wesentlich zu verbessern.“

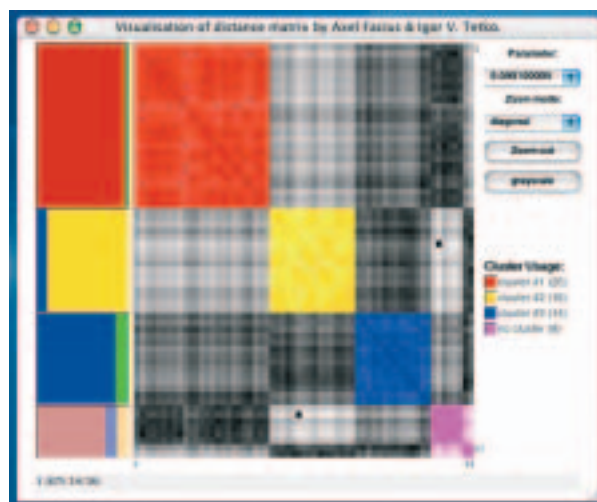
Diese Korrekturen können nun wieder dem Theoretiker Anhaltspunkte für die Verbesserung der Algo-

rithmen geben. Am Institut für Bioinformatik wurden in der Arbeitsgruppe von Klaus Mayer mehr als 10.000 Gene aus dem pflanzlichen Modellorganismus *Arabidopsis* manuell verifiziert und so eine Grundlage für die Datenbank des Genoms geschaffen.

An diesem Beispiel wird auch die duale Aufgabe der Bioinformatiker in Neuherberg deutlich: Ohne ihre Bereitstellung von Daten und Informationen ist eine sinnvolle Arbeit in der molekularen Biologie nicht möglich. Den Erfolg seines Instituts sieht Hans Werner Mewes damit weniger in brillanten Erkenntnissen als in dem hohen praktischen Nutzen, den es für die experimentell arbeitenden Biologen bringt. Der traditionelle Weg wissenschaftlicher Informationsvermittlung in Form von Publikationen,

Bioinformatik - Fruchtbare Vernetzung von Disziplinen

Die Vorstellung darüber, was Bioinformatik bedeutet, geht selbst unter den Fachleuten auseinander. Als interdisziplinäres Gebiet hat sie es mit unterschiedlichen Erbstücken der Disziplinen zu tun, die ihr den Namen gegeben haben. Die Generation der Pioniere ist immer noch den Traditionen der Biologie, der Informatik und der Mathematik verhaftet. Während Biologen oft mit einer gut begründeten Hypothese zur Erklärung ihrer experimentellen Daten zufrieden sind, suchen Informatiker nach pragmatischen Lösungen für komplexe Probleme. Die Lieblingsherausforderung der Mathematiker liegt dagegen im Beweis für deren Lösung, vergleichbar mit Bergsteigern, die sich am Schwierigkeitsgrad, nicht an der Höhe des Berges messen. Jede dieser Auffassungen ist gerechtfertigt, doch erst ihre Kombination führt zum perfekten Ergebnis, das jedoch oft nur durch heuristische, also nicht theoretisch solid begründbare, Methoden erreicht werden kann. Eine Definition der Bioinformatik, die ihre pragmatischen Aspekte in den Vordergrund stellt, bezeichnet die Bioinformatik als Disziplin, die mit den Methoden der Informatik, Statistik und Mathematik versucht, Fragen der molekularen Biologie zu beantworten.

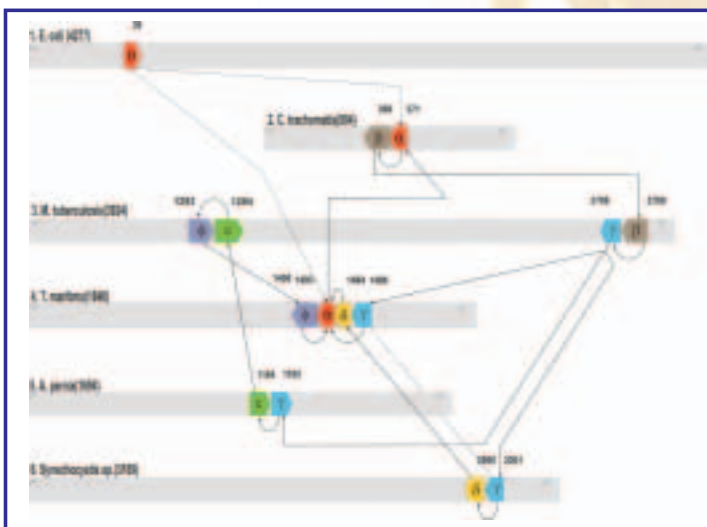


Große Teile

von in Genomen kodierten Proteinen besitzen Merkmale, die auf verwandte biochemische Funktion schließen lassen (sog. Proteinfamilien). Dargestellt ist die Visualisierung einer Proteinfamilienanalyse mittels „Superparamagnetischem Clustering“. Gruppen ähnlicher Proteine sind in farblichen Rechtecken zusammengefasst, ihre Ähnlichkeit zu den anderen Gruppen ist in Grautönen dargestellt.

Bakterielle Genome

sind in sogenannten Operons organisiert. Gene, die biochemisch zusammenwirken, werden oft in Gruppen zusammen abgelesen. Durch den Vergleich verschiedener bakterieller Genome untereinander und die Assoziation von Genen mit bekannter biochemischer Funktion mit solchen unbekannter Funktion kann deren biochemische Rolle abgeleitet werden.



muss durch die Verfügbarkeit strukturierter Informationen in elektronischer Form ergänzt werden. Informationen aus molekularbiologischen Datenbanken werden täglich millionenfach genutzt, allein am GSF-eigenen Server zählen die Bioinformatiker monatlich etwa 25.000 Nutzer.

Der Traum vom Hellsehen

Die andere Seite, komplementär zu dieser notwendigen, aber oft ungeliebten infrastrukturellen Funktion, ist die wissenschaftlich-kreative Aufgabe der Bioinformatik. Erkenntnisse durch die systematische Anwendung der Algorithmen auf große Datenmengen zu gewinnen oder auch durch grundlegende Untersuchungen neue Werkzeuge zu entwickeln, ist verständlicherweise die Lieblingsbeschäftigung der Bioinformatiker. Im Englischen

wird diese Forschung auch gern als „computational biology“ im Gegensatz zur experimentellen Biologie bezeichnet. Im gerade mal pubertären Alter hat die genomorientierte Bioinformatik wie jeder Jugendliche eine Menge von Visionen, die sicher zum Teil Träume bleiben werden.

Dazu gehören klar definierte und zum Teil schon lange offene Fragen und Problemstellungen, deren Lösung aber noch immer auf sich warten lässt. Am bekanntesten dürfte das Faltungsproblem sein. Während die Natur mit großer Zuverlässigkeit die Primärsequenz eines Proteins in seine Raumstruktur überführt, steht die zuverlässige Vorhersage der Raumstruktur aus der Aminosäuresequenz immer noch aus. Erfolgreiche Ansätze beruhen im Wesentlichen darin, bereits bekannte Proteinstrukturen zur Vorhersage zu nutzen. Aber auch die Identifizierung

der genetischen Elemente auf der DNA gelingt längst nicht perfekt, besonders der komplexe Aufbau regulatorischer Bereiche bereitet große Schwierigkeiten. Die Suche nach verbesserten Methoden ist eine Suche nach den richtigen Transformationsalgorithmen, die einzelnen Klassen genetischer Elemente korrekt zu analysieren.

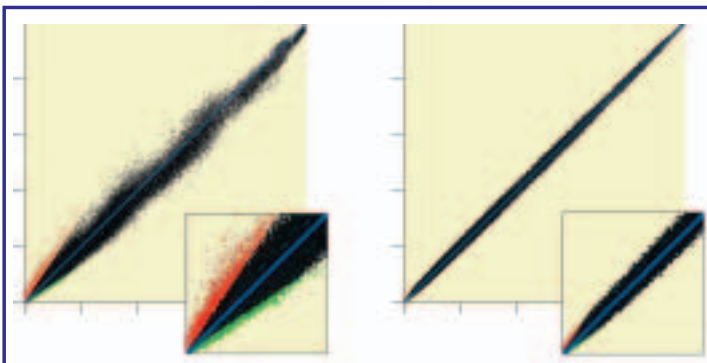
Bereits weit über 100 Genome annotiert

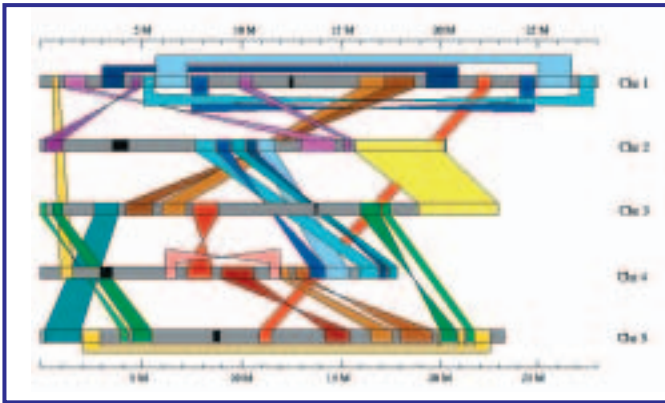
Allein die Tatsache, dass aufgrund der assoziativen Zuweisung zwischen 35 und 70% aller Gene eines Organismus funktionelle Eigenschaften durch Bioinformatikanalyse zugewiesen werden können, beweist die Leistungsfähigkeit der Methode. Durch systematische Anwendung einer großen Zahl von Algorithmen ist es am Institut für Biomathematik gelungen, mehr als 100 Genome automatisch zu annotieren, also Eigenschaften allein durch die Anwendung von Algorithmen und die Nutzung der Sequenz-Struktur-Funktionsbeziehungen vorherzusagen. Die Entwicklung des PEDANT-Systems ist exemplarisch für die Bioinformatik der Genome. Während zunächst manuelle Verfahren Genome detailliert mit großem Aufwand analysiert haben, ist die Flut der erzeugten Daten nur durch die systematische, automatische Anwendung geeigneter Algorithmen zu bewältigen. Die so erzeugten Daten können dann wiederum als Ausgangspunkt für die experimentelle Arbeit dienen, in einem Feedback-Loop kann mit neuen experimentellen Daten die Vorhersagequalität verbessert werden.

Gegenwärtig liefert die Sequenz eines Genoms im günstigsten Fall die Liste aller Einzelteile. Die Interpretation der Daten - und damit die Vision der Bioinformatiker - vergleicht Hans

Genome

Vergleich von Genfamilien mit einem randomisierten Kontroll Datensatz anhand einer statistischen Analyse. Deutlich erkennbar sind die überrepräsentierten (rot) und unterrepräsentierten (grün) Elemente im Testdatensatz (li. Bild). Rot markierte Datenpunkte stellen Kandidaten für funktionelle Elemente, grün markierte Datenpunkte Kandidaten ohne biologische Funktion dar.





Archäologie des Arabidopsis Genoms.

Die fünf Chromosomen des Arabidopsis Genoms sind als graue Horizontalbalken dargestellt. Bereiche mit signifikanter Homologie zu anderen Bereichen im Genom sind mit farbigen Balken markiert. Daraus lässt sich die evolutionäre Geschichte ableiten: Das Arabidopsis Genom hat vor etwa 65 Millionen Jahren eine Genomduplikation durchlaufen. Die daraus entstandenen zwei Kopien wurden unter Restrukturierung und Durchmischung zu der Genomstruktur integriert, wie wir sie heute beobachten.

Werner Mewes mit der Konstruktion eines Autos aus der (auch noch mit schweren Fehlern behafteten) ungeordneten Liste der Einzelteile: „Um das Fahrzeug verstehen zu können, müssen wir zusätzliche Informationen über die Struktur der Teile, ihre Interaktionen (flexibel/statisch) und ihre funktionellen Eigenschaften wie z.B. die Spezifität eines Transportproteins wissen. Erst dann lässt sich eine funktionelle und räumliche Topologie erstellen, die in lebenden Systemen noch erheblich durch die Dynamik des zellulären Lebenszyklus verkompliziert wird.“ Im Gegensatz zum Auto, das zwar altert, aber die Summe seiner Einzelteile und ihre Funktion im Normalzustand konstant erhält, unterliegt aber die Zelle einem ständigen Wandel durch Differenzierung oder Anpassung an den extrazellulären Raum. „Die Summe der mit der zellulären Information verbundenen bekannten Informationen können wir als Biologisches Wissen bezeichnen. Die Darstellung dieses Biologischen Wissens in systematischer, strukturierter Form ist eine unserer wichtigsten Herausforderungen“, resümiert Mewes.

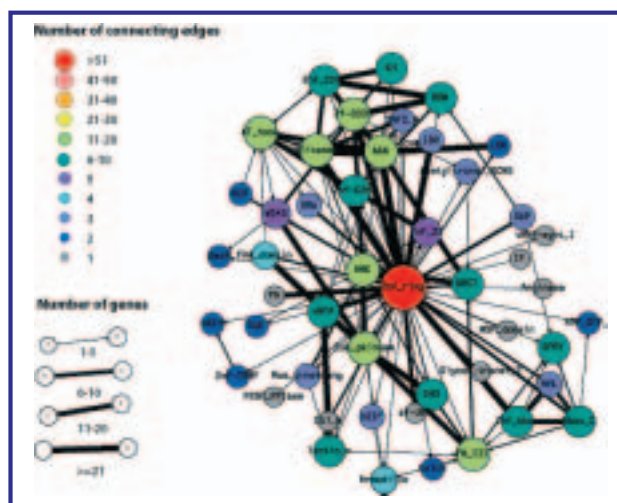
Ordnung in die Datenflut bringen

Um Ordnung zu schaffen, bieten sich funktionelle und strukturelle Klassifikationssysteme an. Da, bedingt durch die evolutionäre Auslese, Struktur- und Funktionsräume im Vergleich zu den kombinatorischen

Möglichkeiten stark limitiert sind, bieten Klassifikationssysteme effiziente hierarchische Möglichkeiten zur Strukturierung von Wissen, ohne im Einzelfall alle Parameter kennen zu müssen. Das am Institut für Bioinformatik entwickelte System des Funktionskatalogs erlaubt die Definition funktioneller Klassen unterschiedlicher Granularität. Während in den 13 Basisklassen nur grobe Unterscheidungen gefunden werden, erlauben die Klassen der unteren Ebenen eine detaillierte Differenzierung. Komplementär zu dieser Einteilung werden funktionelle oder regulatorische Netzwerke, Protein/Protein-Interaktionen, oder Protein/DNA Interaktionen beschrieben. Zur Klassifikation werden Daten aus einer großen Zahl unterschiedlicher experimenteller Ansätze genutzt, so kann die Koexpression von Genen, die in der Expressionsanalyse gefunden werden, wertvolle Hinweise für die funk-

tionelle Zuordnung geben. Um die Vorhersagequalität zu verbessern, arbeiten die Mitarbeiter am Institut an Methoden zur Kombination verschiedener, unabhängiger Parameter, wie z.B. der Expression, der genomischen Topologie und der Protein/Protein-Interaktion.

Ein aktuelles Beispiel für die hohe Bedeutung der Bioinformatik zur Interpretation experimenteller Daten ist die Expressionsanalyse, basierend auf der Kenntnis der genomischen Sequenz. Im Idealfall kann die Expressionsrate, bestimmt durch die Menge der individuellen mRNAs, für jedes einzelne Gen quantifiziert werden. Vor allem die zeitliche Abhängigkeit der Syntheserate unter bestimmten Bedingungen oder die differentielle Änderung in gesundem/kranken Gewebe, erlauben wichtige Aufschlüsse über das dynamische Verhalten, insbesondere über die Koregulation von Gruppen von Genen mit



Biologische Netzwerke

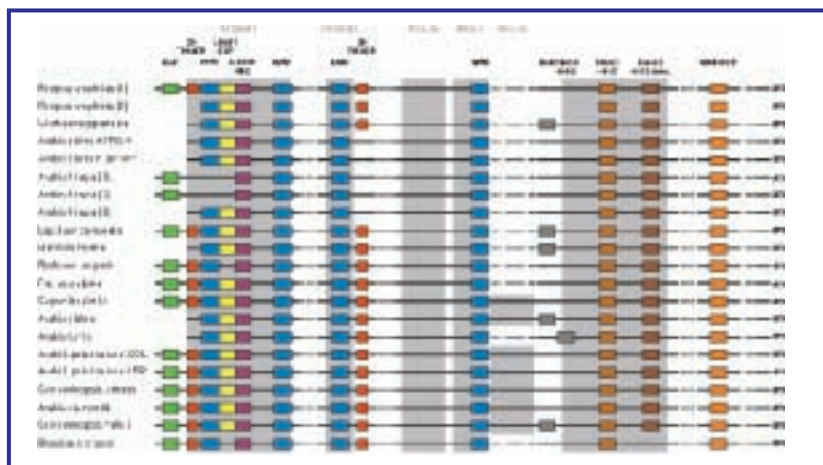
sind hoch komplex. Einzelne funktionelle Bereiche (Domänen) in Proteinen sind oftmals mit anderen solcher Domänen kombiniert. Oft beobachtet man eine Präferenz für nur bestimmte Kombinationen. Dargestellt sind die Kombinationspartner einer Domäne. Deutlich zu sehen ist eine starke Präferenz für nur wenige andere Domärentypen, während die meisten anderen Kombinationen nur selten beobachtet werden.

Brückenschlag zur angewandten Bioinformatik

Mit dem Paradigmenwechsel von der deskriptiven zur systematischen Biologie ist die Bioinformatik zu einem wichtigen Element in der Genomforschung geworden. Ihr Erfolg hängt von der Fähigkeit ab, traditionell und methodisch weit auseinander liegende Disziplinen zu verknüpfen. Im durch das Nationale Genomforschungsnetz (NGFN) geförderten Projekt zur Bioinformatik der Genomanalyse von Säugergenomen (BFAM) versucht das GSF-Institut die Brücke zwischen der experimentellen Genomforschung über die angewandte Bioinformatik zu schlagen und weiter bis hin zur Theorie, die durch Informatiker und Mathematiker vertreten wird. Das Projekt vereint nicht nur ein ganzes Spektrum von Disziplinen, sondern verbindet auch die GSF, die Universitäten und drei Firmen, welche die Bioinformatik kommerzialisieren. Die nächste Generation der Bioinformatiker studiert bereits an den Münchner Universitäten im Rahmen der Bioinformatik Initiative München (BIM). Die Bioinformatik wird in Zukunft weiter an Bedeutung gewinnen. Weniger die zentralen Einrichtungen, sondern vielmehr die kooperativen Interaktionen, die Netzwerke werden die Bioinformatik prägen. Theorie, Methodenentwicklung, technische Implementierung in Form integrierender Werkzeuge werden die experimentellen Methoden der Biologen ergänzen. Die Herausforderung, interdisziplinär denken zu müssen wird auch in Zukunft den Bioinformatikern die Freude an ihrer Arbeit erhalten.

gemeinsamen funktionellen Eigenschaften. Die Bioinformatik muss dabei die Analyse und Interpretation, der meist mehrere tausend Messpunkte umfassenden Expressionsprofile liefern. Matthias Fellenberg hat ein leistungsfähiges System zur Analyse von Funktionsdaten geliefert: „Es ist,“ schwärmt Mewes, „nicht nur in der Lage, zeitliche Verläufe von Expressionsmustern in Gruppen ähnlichen Profils zu clustern, sondern auch die Verknüpfung zu funktionellen Klassen des Funktionskatalogs oder der metabolischen Pfade zu leisten.“

Die nächste Aufgabe liegt im Aufbau eines umfassenden Repositories



Zeitpunkt, Ort und Dauer des Anschaltens einzelner Gene

werden durch regulatorische Elemente gesteuert. Solche Elemente sind sehr klein und sehr schwer zu detektieren. Die Evolution bewirkt jedoch eine natürliche Selektion auf funktionell wichtige regulatorische Elemente. Konservierte Bereiche - die Kandidaten für solche Elemente sind - können mittels „phylogenetic footprinting“ detektiert werden. Dabei werden Gene gleicher Funktion (sog. Orthologe) zwischen Arten verschiedener evolutionärer Distanz (bis zu mehreren Dutzend Millionen Jahren) miteinander verglichen und konservierte Elemente im regulatorischen Bereich als Kandidaten für funktionale Elemente identifiziert.

für Expressionsdaten. Die Mitarbeiter des Instituts versuchen in großen Datenmengen durch Abbildung vor allem auf funktionelle Eigenschaften wie die Kontrolle von Genen durch gemeinsame Promotoren und die Protein/Protein-Wechselwirkungen, Gemeinsamkeiten zu entdecken, die dann als Hypothesen experimentell verifiziert werden können.

Das Ziel – vom Modell zum Menschen

Ein Blick in die Zukunft - Was unterscheidet Gesunde von Kranken? Das Ziel, das sich das Institut für Bioinformatik in den nächsten Jahren gesetzt hat, ist die Nutzung von Genomdaten zur vergleichenden Genomanalyse in Modellsystemen. Was haben Mikroorganismen, Pilze als niedere Eukaryonten, Pflanzen und Säuger gemeinsam und wodurch unterscheiden sie sich voneinander? Dazu wagt Hans Werner Mewes einen Blick in Zukunft: „Vor allem unter der Voraussetzung, dass wir auch in naher Zukunft funktionelle Eigenschaften

nur sehr eingeschränkt vorhersagen können, wollen wir untersuchen, inwieweit wir möglichst mit quantifizierbarer Vorhersagequalität, Eigenschaften von einem System ins andere übertragen können.“ Treffen funktionelle Zuordnungen von einem Bakterium zum anderen auch für höhere Systeme zu? Worin liegen die Details der Mechanismen, in denen sich pathogene von nicht-pathogenen Bakterien unterscheiden? Und, von besonderer Bedeutung, die Frage: Was unterscheidet den gesunden vom pathologischen Status im menschlichen Organismus oder im Modell des Mausgenoms? Welche Gene bzw. ihre Produkte sind in den frühen Stadien der Krankheitsgenese involviert, welche Netzwerke sind betroffen und welche Gene sind an der Antwort des Systems beteiligt? Mewes: „Hier bietet sich die einmalige Chance der intensiven Zusammenarbeit der an der Genomforschung beteiligten Institute und damit die ideale Voraussetzung zur interdisziplinären Zusammenarbeit im Sinne einer echten „win-win“ Situation.“

