Messungen und Datenanalyse in Anbetracht von Rauschen und komplexem Bias – Fortschritte aus verbesserten bioinformatischen Algorithmen*

Paweł P. Łabaj
Chair of Bioinformatics, Boku University Vienna
pawel.labaj@boku.ac.at

Abstract: Die hier vorgestellte kumulative Dissertation im Bereich der Bioinformatik spiegelt die Bandbreite dieses Gebiets wieder und spannt einen Bogen von der direkten Auswertung von Meßergebnissen, über die Detektion von Signalen in einem komplexen Hintergrund, bis zur Frage der optimalen dynamischen Ressourceallokation bei der Exekution mehrstufiger Analyseabläufe. In jedem der Bereiche werden beispielhaft Fortschritte durch moderne bioinformatische Ansätze präsentiert. Insbesondere zeige ich (1) wie Genexpressionsmessungen aus Sequenziermethoden der nächsten Generation verbessert werden können, (2) wie empirische anwendungsspezifische Hintergrundmodelle die Entdeckung neuer funktionaler Proteinsequenzmotive auch in Proteinregionen von starkem kompositionellen Bias erlauben, und (3) erste Schritte auf bei der Überwachung von Ressourcebedarf in heterogenen Analyseabläufen in der Cloud.

1 Einleitung

Bioinformatik ist eine sehr heterogene Forschungsdisziplin: computergestützte Methoden werden eingesetzt, um Erkenntnisse aus biomedizinischen Daten zu gewinnen [Mus11]. Zu diesem Zweck werden entweder etablierte Methoden und Werkzeuge verwendet, um neue Daten bezüglich bestimmter biologischer Fragestellungen zu durchsuchen, oder es werden neue Methoden und Werkzeuge entwickelt und mit gut charakterisierten Datensets getestet. In beiden Fällen ist das Ziel, interessante Muster in Daten aus Laborexperimenten zu finden, die entweder von einem selber oder von Kooperationspartnern gemessen wurden, oder aus öffentlichen Datenbanken stammen.

2 Charakterisierung und Reduktion von Meßrauschen

Im Allgemeinen bestimmt die Meßgenauigkeit von Experimenten die Fähigkeit einer jeden Analyse, relevante Signale zuverlässig zu bestimmen – wie etwa die Unterschiede

*Englischer Titel der Dissertation: 'Measurement and data analysis in the face of noise and complex backgrounds – Advances from improved bioinformatics algorithms'

in einer Rastersuche (‘screen’) für differentielle Expression. Die Charakterisierung und Reduzierung des Rauschens ist daher ein wesentliches Ziel, um so eine höhere Analysenempfindlichkeit zu erreichen. Dies ist besonders für die neuartigen Hochdurchsatztechnologien aktuell, wie beispielsweise Plattformen für die genomweite Expressionsprofilierung mittels Microarray-Hybridisierung oder Sequenziermethoden der nächsten Generation („RNA-Seq“).

In RNA-Seq Protokollen werden Gentranskripte („RNAs“) sequenziert, welche die Aktivität der jeweiligen Gene indirekt widerspiegeln. Die resultierenden Daten bestehen dann aus vielen Millionen kurzer Sequenzstücke(den ‘Reads’). Die Anzahl von Reads, die zu einem bestimmten Gen passen, ist ein Maß seiner Expression (Aktivität). Während die Detektion neuer Transkripte mit der Gesamtzahl der sequenzierten Reads wächst, so stellen die nicht-zufällige Natur der biologischen Sequenzen und die stark asymmetrischen Verteilung deren Häufigkeiten eine wesentliche Herausforderung dar. Das betrifft speziell Studien zur Genregulation, weil die hier biologisch interessanten Gene, die Transkriptionsfaktoren, bereits in sehr geringer Kopienzahl biologisch kausal aktiv sein können.

Beim Einsatz von RNA-Seq zur Quantifizierung von Genexpression stellt sich die Frage, inwieweit Probleme bei der Detektion von Genen in kleinen Kopiezahlen die Qualität der Meßergebnisse bestimmen [JW09, TWP+10]. Meine Arbeit präsentiert dazu die erste umfassende Studie zur Zuverlässigkeit dieser Messungen [LLL+11]: Trotz der generell guten Korrelation zwischen wiederholten Messungen, kann diese Korrelation durch eine kleine Anzahl hoch exprimierter Transkripten dominiert werden, was im Einklang mit der beobachteten geringen Meßgenauigkeit für schwach exprimierte Transkripte steht [MWM+08].

Viele Gene haben verschiedene Spleißformen, die auch biologisch unterschiedliche Aktivitäten zeigen. Die Identifizierung der Spleißstellen ist essentiell für die zuverlässige Bestimmung und Quantifizierung von Spleißformen. Ich habe nun einen hybriden Ansatz für die Analyse von RNA-Seq Reads entwickelt, welcher Wissen über bekannte Spleißformen schon im Alignierungsschritt verwendet, wodurch die Identifizierung von Spleißstellen um das 2–3-fache verbessert werden konnte [LLL+11]. Als Ergebnis konnte ich die Anzahl der zuverlässig gemessenen Transkripte um 50% erhöhen. Um die Methode in der biomedizinischen Forschergemeinschaft leichter zugänglich zu machen, habe ich den Algorithmus auch als bedienerfreundliches Programm implementiert und validiert [LLWK12].

Selbst mit den verbesserten Analysemethoden konnten jedoch nur 41% aller bekannter Transkripte zuverlässig gemessen werden. Meine Analyse zeigt, daß über 75% aller Read-Treffer aus nur 7% des bekannten Transkriptoms kommen, und 99,5% aller Read-Treffer von 41% der bekannten Transkripte herrühren. Folglich stehen nur 0,5% der zugeordneten Reads für die Bewertung der verbleibenden 59% der Transkripte zur Verfügung. Das bedeutet auch, daß der größte Teil der Teststärke der Messung für die Mehrzahl der Transkripte nicht zur Verfügung steht.

Ich habe dann gezeigt, daß der beschriebene Effekt nicht leicht durch eine Erhöhung der Sequenzierungstiefe zu überwinden ist. Jede Verdopplung der Sequenzierungstiefe ergibt nur 5% mehr zuverlässig quantifizierbare Transkripte. Das legt nahe, die Stärke von RNA-Seq zur Identifizierung neuer Transkriptsequenzen einzusetzen, und dann zur effizienten quantitativen Messung benutzerdefinierte Microarrays einzusetzen [LTS+08].

Es sollte angemerkt werden, daß meine Arbeit zur Charakterisierung und Reduktion von Rauschen in RNA-Seq die Notwendigkeit und den Nutzen solcher grundlegender Studien neuartiger Technologien unterstreicht. Komplementär zu Untersuchungen systematischer Fehler, erlaubt uns doch nur ein gutes Verständnis der Rauschcharakteristik einer Meßmethode, effiziente Analyseansätze zu entwickeln und zu validieren, welche zur verläßlichen Extraktion biologisch relevanter Ergebnisse unerläßlich sind.

3 Erkennung von Signalen in einem komplexen Hintergrund

Die funktionale Interpretation genomweiter Daten aus Hochdurchsatz-Technologien bleibt eine wesentliche Herausforderung. Dabei werden weniger der technische Aufwand als konzeptuelle Schwierigkeiten in der Interpretation von Ergebnissen zu wesentlichen Stolpersteinen. Im Allgemeinen ist die Identifizierung biologisch relevanter Unterschiede oder konservierter Muster in einem „Hintergrund“ funktional bedeutungsloser Variation von Interesse. In der Analyse von Proteinsequenzen bedeutet dies die Detektion funktionaler Motive innerhalb nicht-funktionaler Variation. Es wird immer offensichtlicher [AWG+05], daß dieser Hintergrund nicht einheitlich ist, was die Entwicklung adaptiver Algorithmen mit komplexen Modellen zu einem aktiven und wichtigen Forschungsfeld macht.

Für eine Analyse, die eine sensitive und spezifische Detektion biologisch relevanter Signalen in einem variablen Hintergrund ermöglicht, sind nicht-triviale Ansätze erforderlich, die diese Hintergrundeffekte „abziehen“ können [XK05, TST+08]. Es stellt sich heraus, daß die Qualität der Analyse direkt von der Qualität des Hintergrundmodells abhängt [AWZY10]. Einige der ersten solcher kontextspezifischer Ansätze sind die Algorithmen, welche für die Erstellung der Proteinsequenz-Datenbank ‘Pfam’ verwendet wurden. Dabei werden die Hintergrundmodelle so angepaßt, daß sie die Natur der untersuchten Datensätze und die Fragestellung berücksichtigen [SED97]. Analog kann man adaptive Sequenzähnlichkeiten berechnen, indem Substitutionsfrequenzen auf die Aminosäurezusammensetzung jedes einzelnen Proteins eingehen [AWG+05]. Die jüngsten Entwicklungen in dem Feld sind problemspezifische Hintergrundmodelle, wie etwa die Erweiterung des Hintergrundmodells für die Bewertung von multiplen Sequenzalignierungen um die Sekundärstruktur der enthaltenen Sequenzen [SG08a].

Es lohnt sich zu überlegen, wie die Wahl eines Referenzdatensatzes die Fragenstellung beeinflußt, welche in einer Analyse untersucht werden können. Umgekehrt legt die Fragestellung fest, welche Referenzdatensätze für die Konstruktion eines Hintergrundmodells geeignet sind. Speziell, wenn es sich um Regionen mit ungewöhnlich niedriger Komplexität handelt: Diese Regionen sind in der Regel nicht konserviert, und werden daher als funktional irrelevant eingestuft. Also werden diese Bereiche in typischen Sequenzanalysen, wie Homologiesuchen und funktionaler Annotation, von vornherein ausgefiltert [WF93]. Es gibt jedoch zunehmend anekdotische Hinweise darauf, daß auch diese Regionen funktionale Rollen haben [KO03, KH06], wobei die prominentesten Beispiele homopolymere Aminosäureketten sind, die mit einer Reihe von Krankheiten in Verbindung gebracht werden [SG08b].

Bemerkenswert ist, daß die beobachteten Häufigkeiten dieser Ketten sich nicht nur der

Analyse mit Standardmodellen widersetzen [KBB+02]. Ich konnte zeigen, daß sie selbst von Markov-Ketten höherer Ordnung nicht vorhergesagt werden können [LSK11]. In der Arbeit, die ich hier präsentiere [LLB+10, LSK11], wird der Stellenwert der Entwicklung und Validierung von Modellen speziell für die gegebene Problemstellung durch eine Beispielstudie unterstrichen. In einer quantitativen Analyse der Verteilung homopolymerer Aminosäureketten konnte ich die unerwartete Anreicherung bestimmter Ketten in Signalpeptiden zeigen.

Mein Ergebnis stellt nun die erste Evidenz für eine funktionale Rolle einer ganzen Klasse von Aminosäureketten dar. Insbesondere zeige ich, daß Homopolymere aus Leucin in Signalpeptiden überrepräsentiert sind, nicht aber Ketten aus anderen hydrophoben Aminosäuren. Dieser Effekt ist am deutlichsten in Säugetieren. In menschlichen Proteinen finden sich circa 2/3 aller Leucinketten in Signalpeptiden – und das, obwohl weniger als 1/5 aller Proteine ein Signalpeptid haben. Das mag auch erklären, warum Leucinketten trotz ihrer generellen Toxizität [OKSI05] so häufig auftreten können. Interessanterweise und im Gegensatz zum allgemeinen Trend [SPG06] konnte ich für diese Homopolymere eine stärkere Konservierung zwischen Spezies zeigen als für die angrenzenden Sequenzen im Signalpeptid. Diese divergieren im Lauf der Evolution rascher als das restliche Protein [LXD+09]. Zusammen weist das stark auf eine noch unbekannte funktionale Rolle der Leucinketten hin.

Aus einem anderen Blickwinkel ist bemerkenswert, daß die langen Leucinketten, welche die am weitest verbreiteten Aminosäurehomopolymere darstellen, vor allem in transienten Proteinregionen vorkommen, welche aber im endgültigen Proteinprodukt ja abgespalten werden. Das zeigt, daß eine systematische Analyse von Proteinsequenzen, ohne eine getrennte Behandlung transienter Regionen ein stark verzerrtes Bild geben kann. Im diesem Fall betrifft das immerhin gut 1/5 aller Proteine in Säugern und verwandten Spezies.

Die hier vorgestellte Arbeit ist noch von weiterer Relevanz: Auch wenn die präsentierte beispielhafte Analyse von Aminosäureketten in Signalpeptiden einen sehr speziellen Anwendungsfall darstellt, so ist die Demonstration unseres Ansatzes der Entwicklung und des Einsatz eines empirischen anwendungsspezifischen Hintergrundmodells, direkt angepaßt an die untersuchten Daten einerseits und die zu beantwortende Frage andererseits, viel allgemeiner nützlich. Generell sucht man nämlich in der Mehrzahl der wissenschaftlichen Fragen nach einem Signal oder Muster innerhalb eines unbekannten komplexen Hintergrunds, da die Struktur und Störfaktoren in den Modellresiduen üblicherweise nicht bekannt sind [LS07].

4 Systeme für komplexe Analysen

Moderne Analysen in der Bioinformatik erfordern oft einen komplexen Ablauf mehrerer voneinander abhängiger Programmschritte. Die effiziente Ausführung dieser Schritte kann durch sogenannte Workflow-Systeme unterstützt werden. Die Anforderungen an solche Systeme in der bioinformatischen Forschung bringen jedoch eine Reihe spezifischer technischer und organisatorischer Herausforderungen mit sich, die über die in anderen wissenschaftlichen Disziplinen und allgemeinen Workflow-Systemen hinausgehen

[Kre01, Ste08].

Neben klassischen statischen Notwendigkeiten, wie der eingänglichen Datenkonvertierung, und Bedürfnissen zur Laufzeit, wie bei kontinuierlichen Konsistenzprüfungen, bringt die Arbeit mit wissenschaftlichen Workflows in der Bioinformatik weiters zusätzliche Herausforderungen über die gesamte Lebensdauer eines wissenschaftlichen Projektes. Insbesondere ist es durchaus üblich, daß der exakte Weg einer Analyse nicht im voraus bekannt ist, weil a priori viele Studien explorativ sind, und Daten untersuchen, die nicht nur neue Meßwerte bringen sondern auch unbekannte Strukturen aufweisen.

Effizienz-Erwägungen werden umso wichtiger, wenn Berechnungen teuer sind. Einerseits ist dies der Fall, wenn die moderne fortgeschrittene Modelle eingesetzt werden. Ihre Anwendung kann bereits für kleine Datensätze ressourcenintensiv sein [LSK11]. Auf der anderen Seite wird auch die Durchführung einfacher ad-hoc Ansätze nicht trivial bei der Analyse massiver Datensätze. Dieses Problem wird mit der rasenden Verbreitung von Hochdurchsatztechnologien wie moderner Microarrays und der Sequenziermethoden der nächsten Generation hochaktuell [Pen11]. Das traditionelle Arbeitsmodell im Gebiet, bei dem alle Daten zur lokalen Analyse aus dem Internet geholt werden, trifft immer mehr auf limitierte Bandbreiten. Oft ist die einzige praktische Lösung, Daten auf Festplatten mit der Post zu verschicken, oder Rechnungen auf entfernte Knoten zu verteilen, wo die Daten bereits in einer 'Cloud Computing'-Umgebung vorliegen.

Workflow-Management-Systeme sollten daher die notwendigen Abstraktionen zur Verfügung stellen, die die effektive Nutzung von Rechen- und Daten-Ressourcen ermöglichen [Rom08], um es Forschern zu erlauben, sich auf die Durchführung der Datenanalyse und deren Interpretation zu konzentrieren, damit sie effizient Entscheidungen treffen können, wann immer in der Analyse menschliche Weichenstellungen nötig werden. Auch wenn jedes der untersuchten gängigen Systeme manche dieser Fragen anspricht, gibt es kein System, das die nötigen Erfordernisse ausreichend erfüllt, und es ist hier noch ausreichend Raum für weitere Entwicklungen. Dabei ist festzuhalten, daß bestehende Workflow-Management-Systeme im Allgemeinen für die wiederholte Anwendung von Standard-Analysen entwickelt worden sind, welche aber typischerweise in einem industriellen Kontext zu finden sind, im Gegensatz zu hochdynamischen akademischen Anwendungsszenarios. Die Spezifikation neuer Komponenten oder Änderungen an einem Workflow sind nicht ausreichend leicht und können nicht flexibel genug auf dynamische Änderungen eingehen, die häufig in typischen wissenschaftlichen Analyse-Workflows in allen Bereichen zu erwarten sind, sei das nun in den Analyse-Anforderungen selbst, in externem Programmcode, oder öffentlichen Datenquellen.

Darüber hinaus setzen viele moderne Workflow-Systeme auf Web-basierte Dienste. Während der Einsatz von Analyse-Code über das Internet eine elegante und einfache Möglichkeit ist, um eine Vielzahl von Analyse-Werkzeugen zu verbinden, so bringen Abhängigkeiten von verteilten Ressourcen zusätzliche organisatorische und technische Herausforderungen. Insbesondere können Serviceverfügbarkeit variieren und Dienstleistungen in ihrer Implementation von Anbietern auch kurzfristig verändert werden [LGG11]. Das kann dann Eingabe- wie Ausgabeformate, Analyseparameter und Semantik betreffen. Selbst wenn Ausfälle und Versionsänderungen rechtzeitig angekündigt werden, so bleiben sie dennoch außerhalb der Kontrolle der Service-Nutzer und des Workflow-Systems. Kri-

tisch ist hier insbesondere, daß solche Änderungen unerwartete Effekte haben können, die im besten Fall zum Analyseabbruch und im schlimmsten Fall falsche oder inkonsistente Ergebnisse. Sogar wenn sogenannte Service Level Agreements eingesetzt werden, bleiben ernsthafte Probleme. Kommunikation ist oft nicht verschlüsselt, und ohne starke Authentifizierung und Datenverschlüsselung können weder Datenintegrität noch Vertraulichkeit garantiert werden. Schließlich sind Web-basierte Dienste auf einzelnen Server-Rechnern für die Analyse neu generierter großen Datensätze aufgrund der schieren Größe der Eingangsdaten und des entsprechend hohen Rechenaufwands begrenzt einsetzbar [Pen11].

Wenn wir uns nun Gedanken zum Ressourcenbedarf machen, so ist es erwähnenswert, daß die verschiedenen Analyseschritte sehr unterschiedliche Anforderungen in Bezug auf Rechenleistung, Speicherverbrauch, Speichersystem-Leistung und -Kapazität haben. Darüber hinaus können sich die Ressourcenbedürnisse einer komplexen Analysewährend ihrer Ausführung ändern. Rechenumgebungen, die eine flexible Zuweisung von Ressourcen erlauben können inzwischen durch die Integration einer Infrastruktur für verteiltes Rechnen geschaffen werden, wobei Cloud-Umgebungen, Rechnernetzwerke ('Compute Grids') und lokalen Rechnergruppen ('Cluster') zum Einsatz kommen [Ste08]. Es zeigt sich, daß das gezielte Management von Rechenressourcen eine weitere Anforderung für ein effizientes Workflow-System darstellt, das groß angelegte wissenschaftliche Analysen in der Bioinformatik unterstützen kann. Aktuelle Implementierungen integrieren jedoch solch ein Ressourcenmanagement nicht. Damit ist eine verteilte Rechen-Infrastruktur noch nicht effizient einsetzbar, wenn eine komplett erfolgreiche Exekution komplexer Analysen gewährleistet werden muß. Ein notwendiger erster Schritt dazu ist die Fähigkeit, die Workflow-Ausführung zu überwachen, um den Ressourcenverbrauchs zu charakterisieren, und den Status verfügbaren Ressourcen zu überwachen. Das ermöglicht dann in weiteren Schritten eine kontinuierliche Optimierung der Ausführung mehrerer Workflows unter Berücksichtigung der Bedürfnisse der gleichzeitig laufenden Programme. Wir haben dafür eine neue Technologie zur Überwachung von Cloud-Ressourcen zusammen mit einer Wissensmanagement-Strategie für die Optimierung von Workflow-Anwendungen Ausführung in Cloud-Umgebungen entwickelt und getestet [ELM+11]. Insbesondere konnten wir zeigen, wie wir Ressourcen-Engpässe zur Laufzeit erkennen und aktives Wissensmanagement einsetzen können, um dynamisch mehr Ressourcen in solchen Fällen zur Verfügung stellen können, um die erfolgreiche komplette Ausführung der Workflow-Anwendung zu garantieren. In zukünftigen Arbeiten wollen wir diese Ansätze in Prototyp-Workflow-Systemen integrieren, um die Konstruktion tiefer, komplexerer bioinformatischer Analysen zu unterstützen, inklusive der semi-interaktiven explorativen Datenanalyse Schritt-für-Schritt, die für zielgerichtete Analysen gerade in unbekanntem Terrain notwendig ist.

5 Publizierte Resultate meiner Arbeit im Überblick

Schlußendlich basieren alle Analysen auf Daten, die an irgendeinem Punkt gemessen worden sind. Dabei unterliegt jede Messung Fehlern, die entweder rein zufällig sind oder einen systemischen Bias widerspiegeln. Während die Notwendigkeit einer Charakterisie-

rung und Entfernung von Bias zunehmend erkannt wird, werden zufällige Fehler (das „ Rauschen“) nach wie vor oft nur implizit behandelt. Ich habe in meiner Arbeit gezeigt, wie Datenverarbeitungsmethoden einen wesentlichen Einfluß auf das Meßrauschen haben, und wie man entsprechend das Rauschen wesentlich reduzieren kann, um so die Sensitivität und Teststärke folgender Analysen quantitativer RNA-Seq Genexpressionsprofile zu erhöhen [LLL+11, LLWK12].

Zur Identifizierung biologisch relevanter Signale in einer See bedeutungsloser Variationen, ist es zuallererst nötig, die Natur dieses Hintergrunds zu verstehen und zu modellieren. Anhand einer beispielhaften Studie der Überrepräsentation bestimmter Aminosäureketten in Proteinsequenzen habe ich einerseits gezeigt, wie sehr applikationsspezifische Hintergrundmodelle in der Analyse von Sequenzen mit kompositionellem Bias notwendig sind. Andererseits demonstriert diese Arbeit und die dabei entdeckte Alleinstellung der in Säugetieren konservierten Überrepräsentation von Leucinketten in Signalpeptiden jedoch auch, daß selbst in schwierigen Situationen mit entsprechenden Modellen quantitative Analysen möglich sind, und zu biologisch besonders spannenden Ergebnissen führen können [LLB+10, LSK11].

Zur Interpretation von Analyseergebnissen sind oft komplexe Abläufe mit mehreren von einander abhängigen Schritten nötig. Dabei werden eine Vielzahl spezialisierter Analyseprogramme und unterschiedlicher Datentypen aus verschiedenen Quellen gemeinsam eingesetzt. Die Exekution dieser Schritte kann durch sogenannte ‘Workflow Systeme’ unterstützt werden. Solche Systeme werden mit den rasch anwachsenden Datenvolumen, die eine interaktive manuelle Analyse erschweren, inzwischen hochaktuell. Insbesondere die sich inzwischen stark etablierten Hochdurchsatzmethoden in Biologie und Medizin tragen dazu bei, wozu auch die in meiner Arbeit untersuchte Genexpressionsprofilierung gehört. Bei der Abwicklung dieser Analysen ist eine effiziente Nutzung der verfügbaren Computerressourcen eine knifflige Herausforderung, weil die einzelnen Analyseschritte so unterschiedliche Anforderungen haben können. Anhand eines typischen Beispiels habe ich mit Kollegen aus der Angewandten Informatik gezeigt, wie man ein System konstruieren kann, das den dynamischen Ressourcebedarf eines laufenden Workflow mit dem Ziel überwacht, letztendlich eine optimale Ressourcenzuteilung in einem ‘Cloud Computing System’ zu erwirken [ELM+11].

6 Zusammenfassung

In der Bioinformatik werden biologische Daten mit computergestützten Methoden analysiert. Das Ziel sind neue Erkenntnisse in den Lebenswissenschaften. Diese Dissertation präsentiert eigenständige Arbeiten aus drei ineinandergreifenden Bereichen des Gebiets.

Die moderne Molekularbiologie verwendet ein immer größeres Repertoire an Instrumenten für genomweite quantitative Messungen. Meßfehler setzen sich stets aus zufälligem Rauschen und systematischen Abweichungen zusammen. Meine Untersuchung einer neuen Plattform zur Profilierung von Genexpression durch Sequenziermethoden der nächsten Generation („RNA-Seq “) ergänzt Studien, die Quellen systemischer Fehler identifizieren. In meiner Arbeit zeige ich nämlich, wie Meßrauschen durch die Weiterverarbeitung der

Daten beeinflußt wird, und wie dieser Fehler beträchtlich reduziert werden kann und so die Sensitivität und Effizienz nachfolgender Analysen deutlich erhöht.

Zusätzlich zu der verbesserten genomweiten Messung quantitativer Kenngrößen, kann man auch ein exponentielles Wachstum biologischer Sequenzdaten beobachten. Damit bekommt die funktionale Interpretation dieser Sequenzen eine Schlüsselfunktion in der Bioinformatik. Dabei ist die Suche von biologisch relevanten typischen Mustern und Unterschieden vor dem Hintergrund einer funktional bedeutungslosen Varianz von Interesse. Anhand einer Studie des unerwartet häufigen Auftretens von Homopolymeren bestimmter Aminosäuren zeige ich die Notwendigkeit anwendungsspezifischer Hintergrundmodelle für eine Untersuchung von Sequenzregionen mit unüblichen Aminosäurezusammensetzungen.

Mit der Vielzahl an Datenquellen wachsen schließlich Komplexität und Rechenaufwand einer typischen Analyse, und legen Systeme zur Steuerung des Analysenablaufs nahe. Ich zeige hier, wie ein solches System aufgebaut werden kann, das dynamisch die heterogenen Ressourcenbedürfnisse der einzelnen Analyseschritte überwacht, um die Effizienz in einer Rechnerwolke zu optimieren.

Literatur

[AWG+05] Stephen F. Altschul, John C. Wootton, E. Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schäffer und Yi-Kuo Yu. Protein database searches using compositionally adjusted substitution matrices. *FEBS J.*, 272(20):5101–9, 2005.

[AWZY10] Stephen F. Altschul, John C. Wootton, Elena Zaslavsky und Yi-Kuo Yu. The Construction and Use of Log-Odds Substitution Scores for Multiple Sequence Alignment. *PLoS Comput Biol*, 6(7):e1000852, 07 2010.

[ELM+11] Vincent Chimaobi Emeakaroha, Paweł Łabaj, Michael Maurer, Ivona Brandic und David P. Kreil. Optimizing Bioinformatics Workflows for Data Analysis Using Cloud Management Techniques. In *The 6th Workshop on Workflows in Support of Large-Scale Science*, Seattle, November 2011.

[GGM+10] Malachi. Griffith, Obi L. Griffith, Jill Mwenifumbo, Rodrigo Goya, A. Sorana Morrissy, Ryan D. Morin, Richard Corbett, Michelle J. Tang, Ying-Chen Hou, Trevor J. Pugh, Gordon Robertson, Suganthi Chittaranjan, Adrian Ally, Jennifer K. Asano, Susanna Y. Chan, Haiyan I. Li, Helen McDonald, Kevin Teague, Yongjin Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B. Morin, Steven J. M. Jones, Isabella T. Tai und Marco A. Marra. Alternative expression analysis by RNA sequencing. *Nat Methods*, 7(10):843–847, 2010.

[JW09] Hui Jiang und Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics (Oxford, England)*, 25(8):1026–1032, April 2009.

[KBB+02] Samuel Karlin, Luciano Brocchieri, Aviv Bergman, Jan Mrazek und Andrew J. Gentles. Amino acid runs in eukaryotic proteomes and disease associations. *Proc Natl Acad Sci U S A*, 99(1):333–8, 2002.

[KH06] Igor B. Kuznetsov und Seungwoo Hwang. A novel sensitive method for the detection of user-defined compositional bias in biological sequences. *Bioinformatics*, 22(9):1055–1063, 2006.

[KO03] David P. Kreil und Christos A. Ouzounis. Comparison of sequence masking algorithms and the detection of biased protein sequence regions. *Bioinformatics*, 19(13):1672–81, 2003.

[Kre01] David P. Kreil. *From General Scientific Workflows to Specific Sequence Analysis Applications: The study of compositionally biased proteins*. Dissertation, University of Cambridge, 2001.

[LBA+09] Joshua Z. Levin, Michael F. Berger, Xian Adiconis, Peter Rogov, Alexandre Melnikov, Timothy Fennell, Chad Nusbaum, Levi A. Garraway und Andreas Gnirke. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol*, 10(10):R115, 2009.

[LGG11] Burkhard Linke, Robert Giegerich und Alexander Goesmann. Conveyor: a workflow engine for bioinformatic analyses. *Bioinformatics*, 27(7):903–911, 2011.

[LLB+10] Paweł P. Łabaj, Germán G. Leparc, Anaïs F. Bardet, Günther Kreil und David P. Kreil. Single amino acid repeats in signal peptides. *FEBS Journal*, 277(15):3147–3157, 2010.

[LLL+11] Paweł P. Łabaj, Germán G. Leparc, Bryan E. Linggi, Lye Meng Markillie, H. Steven Wiley und David P. Kreil. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, 27(13):i383–i391, 2011.

[LLWK12] Paweł P. Łabaj, Bryan E. Linggi, H. Steven Wiley und David P. Kreil. Improviong RNA-Seq precision with MapAl. *Frontiers in Genetics*, 2012. in press.

[LS07] Jeffrey T Leek und John D Storey. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet*, 3(9):e161, 09 2007.

[LSK11] Paweł P. Łabaj, Peter Sykacek und David P. Kreil. An analysis of single amino acid repeats as use case for application specific background models. *BMC Bioinformatics*, 12(1):173, 2011.

[LTS+08] Germán G. Leparc, Thomas Tüchler, Gerald Striedner, Karl Bayer, Peter Sykacek, Ivo L Hofacker und David P Kreil. Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Res*, Dec 2008.

[LXD+09] Yu-Dong Li, Zhong-Yu Xie, Yi-Ling Du, Zhan Zhou, Xu-Ming Mao, Long-Xian Lv und Yong-Quan Li. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene*, 436(1-2):8–11, 2009.

[Mus11] Arcady Mushegian. Grand Challenges in Bioinformatics and Computational Biology. *Frontiers in Genetics*, 2(0), 2011.

[MWM+08] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer und Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628, Juli 2008.

[OKSI05] Yoko Oma, Yoshihiro Kino, Noboru Sasagawa und Shoichi Ishiura. Comparative analysis of the cytotoxicity of homopolymeric amino acids. *Biochim Biophys Acta*, 1748(2):174–9, 2005.

[Pen11] Elizabeth Pennisi. Will Computers Crash Genomics? *Science*, 331(6018):666–668, 2011.

[Rom08] Paolo Romano. Automation of in-silico data analysis processes through workflow management systems. *Briefings in Bioinformatics*, 9(1):57–68, 2008.

[SED97] Erik L. L. Sonnhammer, Sean R. Eddy und Richard Durbin. PFAM: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–20, 1997.

[SG08a] Ruslan I. Sadreyev und Nick V. Grishin. Accurate statistical model of comparison between multiple sequence alignments. *Nucl. Acids Res.*, 36(7):2240–2248, 2008.

[SG08b] Pratibha Siwach und Subramaniam Ganesh. Tandem repeats in human disorders: mechanisms and evolution. *Front Biosci*, 13:4467–84, 2008.

[SPG06] Pratibha Siwach, Saurabh D. Pophaly und Subramaniam Ganesh. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. *Mol Biol Evol*, 23(7):1357–69, 2006.

[Ste08] Lincoln D. Stein. Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nat Rev Genet*, 9(9):678–688, 2008.

[TST+08] Morgane Thomas-Chollier, Olivier Sand, Jean-Valéry Turatsinze, Rekin's Janky, Matthieu Defrance, Eric Vervisch, Sylvain Brohée und Jacques van Helden. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research*, 36(suppl 2):W119–W127, 2008.

[TWP+10] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold und Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, 2010.

[WF93] John C. Wootton und Scott Federhen. Statistics of local complexity in amino-acid-sequences and sequence databas. *Computers & chemistry*, 17(2):149–163, 1993.

[WI05] Zhijin Wu und Rafael A Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, 12(6):882–893, August 2005.

[XK05] Jun Xie und Nak-Kyeong Kim. Bayesian Models and Markov Chain Monte Carlo Methods for Protein Motifs with the Secondary Characteristics. *J Comp Biol*, 12(7):952–970, 2005.

Paweł P. Łabaj wurde am 31. Juli 1982 in Mielec in Polen geboren. Im Jahr 2001 absolvierte er in Zabrze das Gymnasium in einer physikalisch-mathematischen Klasse. In 2006 erhielt von der Schlesischen Technischen Universität in Gliwice den Titel Magister-Ingenieur der Informatik mit Auszeichnung. Seine Masterarbeit über ein System zur automatischen Analyse und Mustererkennung durch neuronale Netze war im Jahr 2006 als beste Arbeit des Jahres in ganz Polen ausgezeichnet worden. Die Arbeit entstand in Zusammenarbeit mit dem Institut für Medizinische Technik und Ausstattung in Zabrze. Danach war Paweł als wissenschaftlichen Mitarbeiter in der Abteilung Biomedizinische Informatik tätig. Im Jahr 2007 nahm er eine Stelle als Predoctoral Research Fellow am Chair of Bioinformatics an der Universität für Bodenkultur Wien an. Seit seiner Promotion ist er als Junior Scientist an diesem Institut beschäftigt. Er interessiert sich für verschiedene Aspekte der Bioinformatik, vor allem die biologische Sequenzanalyse, Next Generation Sequencing, Klassifizierungsmethoden, Merkmalsextraktion, Mustererkennung, Datenintegration und Workflow-Systeme.