Validierung von Genomannotationen

Exposé einer Diplomarbeit

Betreuer: Ulf Leser, Felix Bübl
Bearbeiter: Raphael A. Bauer

28. November 2005

Lage

Um Daten wie Experimentalergebnisse und Annotationen auf einem Genom zu verankern verwendet man in Forschung und Wirtschaft hauptsächlich zwei Verfahren, die sich in Teilen ergänzen.
Bei Verfahren eins (absolute Koordinaten) wird eine direkte Ortsangabe (engl: "location") mittels Chromosomalkoordinaten angewendet [3]. Bei Verfahren zwei (relative Koordinaten) verwendet man eine Referenz zu einer anderen Annotation. Diese Annotation kann ein bestimmtes Gen sein, was sich besonders bei funktionalen Annotationen anbietet. Beide Positionierungssysteme haben das Problem, dass sie bei sich verändernden Basisdaten vollkommen verrutschen können. Daraus resultierend können gravierende Inkonsistenzen entstehen.
Bei der Genome Assembly der Ensembl Datenbank wird etwa jedes halbe Jahr eine neue Version der Daten veröffentlicht, in denen gefundene Fehler bereinigt wurden. Fehler können zum Beispiel korrigierte Basen an bestimmten Stellen sein. Im schlimmsten Falle verschieben sich Chromosomalkoordinaten, wenn durch Nachsequenzieren festgestellt wird, dass es bestimmte Basenpaare nicht gibt, oder neue Basen eingefügt werden müssen. Dies ist besonders bei repetitiven Bereichen im Genom der Fall.
Durch diese Veränderung der Länge und/oder Verschiebung der Abschnitte in der Genomsequenz kann die bisherige Position einer Annotation ungültig werden.
Hat man eine Verankerung auf Genen der Ensembl Datenbank (Gene Assembly) gewählt, ergibt sich das zusätzliche Problem, dass diese Daten meist sogar im Monatsrythmus aktualisiert werden. Verändert nun aufgrund neuer Erkenntnisse ein Gen seine Anfangsposition im Genom, treten Seiteneffekte auf, die die Ergebnisse eines auf diesen Annahmen aufbauenden Experiments verändern, und/oder in einem völlig neuen Licht erscheinen lassen.
Zusammenfassend ergibt sich bei beiden genannten Verfahren folglich das Problem, dass sich die Datenbasis (Genome, Gene Assembly) im schlimmsten Falle jeden Monat ändert und somit Annotationen ungültig werden oder ganz verschwinden. Treten also Fehler bei der Zuordnung von Annotationen auf, müssen davon betroffene Experimentalergebnisse zum einen erkannt, in einem nächsten Schritt hinterfragt und dann gegebenenfalls korrigiert werden.

Ziel

In dieser Diplomarbeit soll eine neue Art der Validierung von Annotationen auf unstabilen Daten untersucht und entwickelt werden. Um das Problem einzukreisen, soll das in diesem Falle auf biologischen Daten der Ensembl Datenbank (Humanes Genom, Humaner Gensatz) und Testdaten einer Berliner Biotechnologiefirma erfolgen.
Das Ziel ist ein prototypisches System, das anhand von zu definierenden Constraints in der Lage ist, automatisch bei einer Änderung der Datenbasis Fehler und Inkonsistenzen zu erkennen, und deren Fehlerklasse zu bestimmen.
Dies schliesst die Entwicklung und Evaluation eines effizienten Algorithmus' mit ein, der es erst praktikabel erscheinen lässt, eine Constraintprüfung auf dem Humanen Genom mit seinen 3 Milliarden Basenpaaren durchzuführen. Des weiteren soll in enger Kooperation mit Biochemikern und Biologen analysiert werden, inwieweit eine automatische Korrektur von bestimmten Fehlerklassen möglich und auch sinnvoll ist.

Maßnahme

Anforderungsanalyse

In der ersten Phase werden in Zusammenarbeit mit Biochemikern und Biologen die domänenspezifischen Constraints identifiziert, die für eine Qualitätssicherung auf dem Genom wichtig und nötig sind. Dabei wird untersucht, inwieweit sich diese Constraints durch eine Ontologiesprache ausdrücken lassen.
Sollten sich im Verlauf der Anforderungsanalyse bereits esistierende Ontologien wie die Sequence Ontolgy [4], oder Allzweckontologiesprachen wie OWL [2] als untauglich erweisen, wird die Verwendung eigener domänenspezifischer Constraints und Regelfälle angestrebt.

Prototypische Implementierung

In einer prototypischen Implementierung auf Basis des Eclipse Frameworks [1] soll die Funktionsfähigkeit der gemachten Annahmen bewiesen werden. Hierfür wird ein effizienter Algorithmus entwickelt und implementiert, der basierend auf dem Constraint Regelsatz performant Annotationen auf ihre Validität hin prüft. Dieses System wird in der Lage sein, Fehler zu erkennen, zu markieren, und den Fehlern entsprechende Maßnahmen zu ergreifen. Um den Systemaufbau unter praxisnahen Bedingungen testen zu können, werden sämtliche Entwicklungen auf echten Annotationsdaten einer Berliner Biotechnologiefirma und dem Humanen Genom und Gensatz der Ensembl Datenbank [3] validiert.

Literatur

[1] Eclipse Framework, http://www.eclipse.org, überprüft am 28. November 2005

[2] Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, 10 February 2004, OWL Web Ontology Language Guide, W3C Recommendation, http://www.w3.org/TR/2004/REC-owl-guide-20040210/, überprüft am 28. November 2005

[3] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinsci, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark and E. Birney, Ensembl 2005, Nucleic Acids Res. 2005 Jan 1;33 Database issue: D447-D453.

[4] K. Eilbeck, S. Lewis, C Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, 29 April 2005, The Sequence Ontology: a tool for the unification of genome annotations, http://genomebiology.com/2005/6/5/R44, überprüft am 28. November 2005