Comparative sequence analysis and association mining in gene regulation

Christoph Dieterich

August 2004

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

1. Betreuer: Prof. Dr. Martin Vingron
2. Betreuer: Prof. Dr. Stefan Mundlos

Datum der Disputation: 04.März 2005

Für Silke, Eva und Joschua

Contents

List of Figures

List of Tables

Abbreviations

Abbreviations in alphabetical order

A, `A`	adenine
bp	base pair(s)
°C	degree celsius
C, `C`	cytosine
cDNA	complementary DNA; DNA synthesized from mRNA by RT
CDS	coding sequence
ChIP	chromatin immunoprecipitation
CNB	conserved noncoding block; suboptimal local alignment
cRNA	complementary RNA
DNA	deoxyribonucleic acid
FDR	false discovery rate
G, `G`	guanine
Kb	kilobase(s); 1,000 nt
LPS	lipopolysaccharide
Mb	megabase(s); 1,000,000 nt
mRNA	messenger RNA
nt	nucleotide(s)
PAM	point accepted mutation
PCR	polymerase chain reaction
POA	partial order alignment
POG	partial order graph
PSSM	position specific score matrix
PWM	position weight matrix
RNA	ribonucleic acid
RT	reverse transcription
RTase	reverse transcriptase; an enzyme
RT-PCR	reverse transcription-polymerase chain reaction
SNP	single nucleotide polymorphism
SQL	structured query language
SRF	serum response factor
T, `T`	thymine
TF	transcription factor
TFBS	transcription factor binding site
tRNA	transfer RNA

U, `U`	uracil
UTR	untranslated region; part of mRNA transcripts

For a compilation of IUPAC symbols for nucleotide nomenclature see Appendix F.

Appendix

A Kurzzusammenfassung

Diese Promotionsarbeit beschäftigt sich mit der Steuerung der Transkription von Genen in Vertebraten im Allgemeinen und Säugern im Speziellen. Der Vorgang der Transkription ist das Abschreiben von Genen in eine RNA Kopie und stellt den ersten Schritt auf dem Weg zum Genprodukt dar. Unmittelbar um den Transkriptionsstart liegen Sequenzelemente, die hauptsächlich für die Effizienz des Transkriptionsvorganges relevant sind. Diese Sequenzelemente werden als Promotoren bezeichnet. Die Ausdehnung von Genen und Bereichen zwischen den Genen ist in Säugetieren beträchtlich groß (bis zu 1 Mb). Transkriptionsstartpunkte und somit Promotoren lassen sich experimentell nur schwer erfassen.

Mein Ansatz nutzt die vergleichende Analyse orthologer Sequenzbereiche aus verschiedenen Spezies. Die Grundidee ist das Aufspüren von konservierten regulatorischen Sequenzelementen und Promotorbereichen. Diese Elemente sind deshalb konserviert weil ein Ausfall oder Fehlen zur Fehlsteuerung des Genprodukts führen würde, das ähnliche Aufgaben in den entsprechenden Spezies verrichtet. Regulatorische Sequenzelemente stehen somit unter selektivem Druck. Mithilfe verschiedener etablierter und neu entwickelter Algorithmen wurde eine umfassende Sammlung solcher konservierter regulatorischer Elemente berechnet. Diese Daten wurden dann gegen bekannte Transkriptionsfaktorbindemuster abgeglichen. Die resultierenden Daten wurden rechnerisch mit den Resultaten anderer biologischer Experimente verknüpft. Aufbauend auf einer statistischen Analyse der Verteilung der relevanten Muster, läßt dies Schlußfolgerungen zur Funktionalität dieser Muster zu.

Die Arbeit behandelt die folgenden Bereiche:

1. **CORG Datenbank** Das CORG (Comparative Regulatory Genomics) Projekt vereint eine Vielzahl an Methoden und Daten zur Vorhersage von Promotorbereichen. Zunächst werden Gruppen homologer Gene für die vergleichende Sequenzanalyse definiert. Upstreambereiche der Gene, die den eigentlichen Promotorbereich mit hoher Wahrscheinlichkeit enthalten, werden auf Sequenzebene miteinander verglichen. Hierzu wird der Waterman-Eggert Algorithmus zum Auffinden lokaler Sequenzähnlichkeiten herangezogen. Um eine Signifikanz der beobachteten Alignmentpunktzahl zu bestimmen wird die Verteilung aus zufälligen Alignments als Poissonverteilung approximiert. Nicht-translatierte Exons werden durch den Vergleich mit assemblierten EST Sequenzen detektiert. Transkriptionsfaktorbindestellen werden innerhalb konservierter Sequenzabschnitte durch

Konsensmuster oder Gewichtsmatrizen vorhergesagt. Alle Daten sind in einer relationalen Datenbank abgelegt, die über eine graphische Benutzeroberfläche und PERL-Schnittstelle zugänglich ist.

2. **Vorhersage von Zellzyklusregulatoren.** Periodisch exprimierte Gene wurden in 5 Gruppen eingeteilt, die den Zellzyklusphasen M/G1, G1/S, S, G2 und G2/M entsprechen. Für jede dieser Gruppen wurde ein Profil der konservierten Bindungsstellen in den dazugehörenden Upstreambereichen erstellt. Die Verteilung der Bindungsstellen über alle Gengruppen wurde in Relation zu einem Nullmodel gesetzt. Das Nullmodell beruht auf der Annahme, dass zufällige Bindungsstellen proportional zur Größe des Sequenzsuchraums auftreten. Die Abweichung von beobachteter Verteilung von Bindungsstellen zu erwarteter Verteilung wurde mit einem exaktem und einem approximativen Test bewertet.

3. **Detailstudie SRF-induzierter Gene.** Der Serum Response Factor (SRF) ist an einer Vielzahl biologischer Prozesse beteiligt: u.a. Immunantwort, Herzentwicklung, embryonale Frühentwicklung und Neurogenese. Im Rahmen dieser Arbeit wurden die Promotoren von zwei Gengruppen studiert, welche durch SRF induziert wurden.

   Die erste Gruppe umfasst Gene, welche durch SRF in *Srf -/-* embryonalen Stammzellen der Maus induziert werden. Die zweite Gruppe beinhaltet Gene die durch Stimulation humaner dendritischer Zellen mit LPS als Teil der Immunantwort induziert werden. Die Rolle von SRF und die Qualität der CORG Promotoranalyse wurden zunächst eingehend in embryonale Stammzellen der Maus studiert. Die signifikante Anreicherung von konservierten SRF Bindungsstellen und deren experimentelle Validierung mittels Chromatin-Immunopräzipitation lassen auf eine hohe Güte der CORG Promotoranalyse schließen.

Das CORG Projekt wurde im Rahmen dieser Promotionsarbeit etabliert und dessen Nutzen anhand von biologischen Fragestellungen klar belegt. Das CORG Rahmenwerk ist offen und flexible gestaltet. Neue Datenströme aus einer Vielzahl an Experimenten werden so noch den Weg in die CORG Architektur finden.

B Erklärung zur Urheberschaft

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Christoph Dieterich Berlin, im August 2004

C List of related publications

Michael M., Dieterich C., Stoye J. (2004) *Suboptimal Local Alignments across Multiple Scoring Schemes.* WABI 2004 - 4th Workshop on Algori thms in Bioinformatics, in press.

Dieterich C., Rahmann S., Vingron M. (2004) *Functional Inference from Non-random Distributions of conserved predicted Transcription Factor Binding Sites.* Bioinformatics. Suppl 1:I109-I115.

Samsonova A., Dieterich C., Vingron M., Brazma A. (2004) *Search for Regulatory Motifs in the Drosophila melanogaster genome.* BGRS 2004 - 4th International Conference on Bioinformatics of Genome Regulation and Structure.

Dieterich C, Herwig R, Vingron M. (2003) *Exploring potential Target Genes of Signaling Pathways by predicting conserved Transcription Factor Binding Sites.* Bioinformatics. Suppl 2:II50-II56.

Dieterich C., Wang H., Rateitschak K., Luz H., Vingron M. (2003) *CORG: A Database for Comparative Regulatory Genomics.* Nucleic Acids Research 31(1):55-7.

Dieterich C., Cusack B., Wang H., Rateitschak K., Krause A., Vingron M. (2002) *Annotating regulatory DNA based on Man-mouse genomic comparison.* Bioinformatics. 18 Suppl2:S84-90.

D Curriculum vitae

Dipl.-Biol. Christoph Dieterich
Leichhardtstraße 59
D-14195 Berlin, Germany
Tel.: ++49 – 30 – 83 33 85 30
E-mail: Christoph.Dieterich@molgen.mpg.de

Date of birth: January 7, 1975
Place of birth: Peine, Germany
Citizenship: German
Marital status: Married, two children

Degrees

MRes in Bioinformatics, University of York, UK. Degree with distinction

- Research experience in EST analysis, exploring the protein fold space and JAVA application development.

MSc in Technical Biology, University of Stuttgart, Germany. Grade: 1.0

- Diploma thesis: "In vitro infection models for *Candida albicans*", Fraunhofer Institute for Interfacial Engineering and Biotechnology, Grade: 1.0
- Research experience in protein engineering, cell biology and prokaryotic genetics.

Education

2001 - present PhD student in the department of Computational Molecular Biology, MPI for Molecular Genetics, Berlin

2000 - 2001 Bioinformatics student at the University of York, UK

1994 - 2000 Biology student at the University of Stuttgart, Germany

1985 - 1994 Staufer-Gymnasium in Waiblingen, Germany (Abitur: 1.3)

Prizes and Awards

2000 - 2001 Scholarship from the Serono Pharmaceutical Research Institute, Geneva, Switzerland.

2000 Hugo Geiger prize for junior scientists (founded by the Bavarian government).

1996 - 2000 Scholarship from the Hanns-Seidel-Stiftung, Munich, Germany.

E Availability

E.1 The modified SIM implementation

The original SIM program is a linear-space implementation of the Waterman-Eggert algorithm by Huang and Miller (1991). We added several features to the program: significance assessment of alignment score (p-value computation), handling of multiple FASTA files and score matrix files.

Our program `SIMSTAT` is available from `http://corg.molgen.mpg.de/software`

E.2 Other software

All software related to the build up, maintenance, visualization and querying of CORG is available on request. Please contact `christoph.dieterich@molgen.mpg.de`.

E.3 CORG Database

MySQL dumps as well as flatfile dumps of the CORG database are made available via `http://corg.molgen.mpg.de/downloads`. Alternatively, a DAS server exists that provides the same data. Visit `http://tomcat.molgen.mpg.de:8080/das` for an overview of the available resources.

F IUPAC nucleotide ambiguity codes

Symbol	Meaning	Nucleic Acid
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
M	A or C	
R	A or G	
W	A or T	
S	C or G	
Y	C or T	
K	G or T	
V	A or C or G	
H	A or C or T	
D	A or G or T	
B	C or G or T	
X	G or A or T or C	
N	G or A or T or C	

Reference: IUPAC-IUB SYMBOLS FOR NUCLEOTIDE NOMENCLATURE: Cornish-Bowden (1985) Nucl. Acids Res. 13: 3021-3030.