遺伝子ネットワークモデル化研究チーム

Biomedical Knowledge Discovery Team

チームリーダー　**SCHÖNBACH, Christian**

遺伝子ネットワークの同定においては，遺伝子構造の種内および種間配列差の比較および分子機能の推定が不可欠である。しかしながら，機能の判明していない遺伝子産物（すなわちタンパク質コードおよび非タンパク質コード転写物）が数多く存在し，調節機構の複雑度が高まるにつれ，この作業は困難になっている。このため，我々はコンピュータによる方法および実験的方法を用いて，調節回路に対する理解を深めることのできるゲノム，トランスクリプトームおよびプロテオームレベルでのパターン関連，選択的スプライシングによって生成する変異，反復要素，選択的翻訳開始点の体系的な同定法を検討する。

1. トランスクリプトームとプロテオームレベルにおけるネットワーク調整のより良い理解のための生物学的知識の発見（Schönbach，Kurochkin，長嶋）

GenBank に登録されたマウス cDNA 配列のゲノムワイドな分析から，CDS を含むゲノムエクソンの長さの平均が非 CDS-ゲノムエクソンの半分であることが確認された。単一のゲノムエクソンは，マルチエクソン遺伝子の選択的スプライシングを受けたゲノムエクソンよりも長い。選択的スプライシングを受けたエクソンの長さの平均は，構成的スプライシングを受けたエクソンの半分以下である。その生物学的意味および重要性を解明するために，他の配列パターンと関連した所見を現在検討している。

翻訳開始は，Kozak コンセンサス配列に隣接する最初の AUG コドンより開始すると考えられている。同時に，Kozak 配列を有する最初の AUG が常に翻訳開始に用いられるわけではないことも分かっている。内側にあるリボソーム結合部位を通じたリークスキャンおよび再開などの機構により，第 2 の AUG コドンにおいて翻訳が開始されることがある。GenBank リソース 141 より抽出した完全コード配列情報を有するマウス cDNA を用いて，1 つの遺伝子から 2 つのタンパク質を生成する選択的スプライシング下での翻訳開始点を検討した。GenBank 由来オープンリーディングフレームと 400 bp 以上オーバーラップする，予測された選択的オープンリーディングフレーム（2,444 ORF）候補をコンピュータによってカテゴリー化し，Kozak コンテクスト配列，AUG の距離，異種間保存，タンパク質ドメイン/モチーフ，膜内外電位差および既知の疾患遺伝子との潜在的類似性についてアノテーションを付与した。この手順により，高い翻訳可能性を示す 1,597 個の選択的 ORF が得られた。特に興味深いものとしては，ヒトタンパク質との相同性を有する候補，およびタンパク質特性の変化（すなわち正常 ORF の膜貫通タンパク質から選択的 ORF の可溶性タンパク質への変化）があった。候補の 1 つは，選択的 ORF においてヒト相同性を有するジンクフィンガータンパク質をコードする。候補タンパク質は組織特異的であると見られ，抗体を用いて実験的に確認されている。これらの遺伝子産物はプロテオームの複雑度を高めるだけでなく，発生過程，成長（すなわち転写活性化/抑制の切り替え）または誘導過程（すなわち免疫応答，新規 T 細胞エピトープ源）の新たな調節機構を意味している。

これらの過程が分からなければ，インスリン抵抗性あるいは肝炎などの病理機構を理解することは難しいであろう。特異的免疫系機能がどのように代謝過程と関連しているかをより良く理解するために，モデルとしてペルオキシソームを選択している。ペルオキシソームは単膜細胞内オルガネラであり，非常に炭素数の多い脂肪酸の β-および α-酸化などの様々な代謝的反応を行う。代謝反応を触媒する酵素はほとんどが判明しており，標的シグナルを通じてペルオキシソームに侵入する。ペルオキシソーム増殖調節因子による転写調節機構以外は，ペルオキシソーム酵素が翻訳後にどう修飾をうけ，分解されるか否か，またはこれらがどのように行われるかほとんど知られていない。従って，これらの過程に関与する候補を同定するためのコンピュータスクリーニング法を開発および適用した。11 種類の遺伝子によりコードされる新規ペルオキシソーム標的の新たな候補 28 種類を同定した。2 つの候補について実験的検証および機能同定を実施中である。ここでも，既知のペルオキシソーム経路と見えたものが調節という点では予想より遙かに複雑であることが判明した。本試験の結果より，実験による検証を行わずに異なる組織または種にまたがって所見を外挿する *in silico* および文献による経路の再構築の取り扱いには注意を要することも示唆される。

Identifying gene structure, intra- and inter-species sequence differences and assigning molecular functions are among the first but critical steps towards any network construction. The task is complicated by a large number of gene products (i.e. protein coding and non-protein coding transcripts) with unknown functions and an increasingly complex picture of regulatory mechanisms. Therefore we use both computational and experimental methods to systematically identify and investigate pattern associations on genome, transcriptome and proteome level, variations generated by alternative splicing, repetitive elements, alternative translation initiation that can increase our understanding of regulatory circuits.

In a large-scale descriptive analysis of mouse cDNAs deposited in GenBank we found that CDS-containing genomic exons are on average half as long as non-CDS genomic exons. Single genomic exons are longer compared

to alternatively spliced genomic exons of multi-exon genes. Alternatively spliced exons are on average less than half as long as constitutively spliced exons. These and other sequence pattern-associated findings are now under investigation to understand their biological meaning and significance.

Translation initiation is supposed to start at the first AUG codon that is flanked by a favorable context sequence called Kozak consenus. In the meantime we know that the first AUG that has a favorable Kozak sequence is not always used for translation initiation. Mechanism such as leaky scanning and re-initiation through an internal ribosome entry site may initiate translation at secondary AUG codon. We examined translation initiation under the aspect of alternative translation (two proteins from one transcript) using mouse cDNAs with complete coding sequence information extracted from GenBank 141. Predicted alternative open-reading frames (2,444 ORFs) candidates that overlap at least over 400 bp with the GenBank-derived ORF were computationally categorized and annotated according to Kozak context, AUG distance, cross-species conservation, protein domains/motifs, transmembrane potential and potential similarity to known disease genes. The procedure yielded 1,597 alternative ORFs that show a high probability of being translated. Of particular interest were candidates with homology to human proteins and changes in protein properties (i.e. transmembrane in normal ORF to soluble protein in alternative ORF). One candidate encodes in the alternative ORF a zincfinger protein which has a human homolog. The candidate protein appears to be tissue specific and has been experimentally confirmed using antibodies. These gene products not only increase the complexity of the proteome but imply new control mechanisms of developmental processes, growth (i.e. transcriptional activation/repression switch) or inducible processes (i.e. immune response, source of new T-cell epitopes).

Without knowledge of these processes it will be difficult to understand pathological mechanisms, for example insulin resistance or liver inflammation. For better understanding how specific immune system functions are linked to metabolic processes we have chosen peroxisomes as a model. Peroxisomes are single-membrane subcellular organelles which host various metabolic reactions including β- and α-oxidation of very long-chain fatty acids. The enzymes catalyzing the metabolic reactions are mostly known and enter the peroxisome through targeting signals. Apart from transcriptional control mechanisms by peroxisome proliferator regulators very little is known whether and how peroxisomal enzymes are processed, post-translationally modified and degraded. Therefore we developed and applied a computational screening method to identify candidates involved in these processes. We identified 28 new peroxisome-targeted candidates encoded by 11 genes. Two candidates are in the process of experimental validation and functional characterization. Again we found that seemingly known peroxisomal pathways are far more complex in terms of regulation than anticipated. Our results also indicate that in silico and literature-based pathway re-constructions that extrapolate findings across different tissues and/or species without experimental validation need to be treated with caution.

Research Subjects

1. Non-canonical transcriptional and translation variation and their association with human diseases-gene pathway
2. Computational identification and functional characterization of peroxisomal protein candidates towards their role in obesity and immune function relationships
3. Role of repetitive elements in modulating gene/gene product functions

Staff

Team Leader
Dr. Christian SCHÖNBACH

Research Scientist
Dr. Igor V. KUROCHKIN

Research Associate
Dr. Takeshi NAGASHIMA

Visiting Members
Mr. Klaus VOSS

Trainees
Dr. Diego G. SILVA (JSPS Fellowship)
Mr. Christoph BOCK (REES Fellowship)
Ms. Marlis HERBERTH (Internship)
Mr. Mohammed S. ALI (Internship)

誌 上 発 表 **Publications**

[雑誌]

(原著論文) ∗印は査読制度がある論文

Shibuya T. and Kurochkin I. V.: "Match chaining algorithms for cDNA mapping", Lect. Notes Comput. Sci. **2812**, 462–475 (2003). ∗

Silva D. G., Schonbach C., Brusic V., Socha L. A., Nagashima T., and Petrovsky N.: "Identification of "pathologs" (disease-related genes) from the RIKEN mouse cDNA dataset using human curation plus FACTS, a new biological information extraction system", BMC Genom. (Web) (http://www.biomedcentral.com/1471-2164/5/28/) **5**, 28 (2004). ∗

Schonbach C.: "From masking repeats to identifying functional repeats in the mouse transcriptome", Brief. Bioinf. **5**, 107–117 (2004). ∗

Schonbach C., Nagashima T., and Konagaya A.: "Textmining in support of knowledge discovery for vaccine development", Methods **34**, 488–495 (2004). ∗

[単行本・Proc.]

(原著論文) ∗印は査読制度がある論文

Nagashima T., Silva D. G., Konagaya A., and Schonbach C.: "Computational identification and exploration of immune-related transcripts in mouse", Immunology

2004: Collection of Free Papers presented at 12th Int. Congr. of Immunology and 4th Ann. Conf. of FOCIS, Montreal, Canada, 2004–7, MEDIMOND S.r.l, Bologna, pp. 391–394 (2004).

Fukuzaki A., Nagashima T., Ide K., Konishi F., Hatakeyama M., Yokoyama S., Kuramitsu S., and Konagaya A.: "Genome-wide functional annotation environment for *Thermus thermophilus* in OBIGrid", Proc. 1st Int. Workshop on Life Science Grid (LSGRID2004), Kanazawa, 2004–5〜6, The Japanese Society for Artificial Intelligence, Kanazawa, pp. 89–98 (2004). ∗

（総　説）

Schonbach C.: "Strategy and planning of bioinformatics experiments", The Practical Bioinformatician, edited by Wong L., World Scientific Publishing, Singapore, pp. 31–34 (2004).

Schonbach C. and Matsuda H.: "Mining new motifs from cDNA sequence data", The Practical Bioinformatician, edited by Wong L., World Scientific Publishing, Singapore, pp. 359–374 (2004).

口　頭　発　表　Oral Presentations

（国際会議等）

Schonbach C.: "Immunoinformatics-driven identification and exploration of immune-related transcripts in mouse and man", Singapore Immunoinformatics Symp., (Institute for Infocomm Research, Singapore), Singapore, Singapore, Mar. (2004).

Hosaka J., Matsumura K., Yoshikawa S., Kurochkin I. V., and Konagaya A.: "PBIE toolkit for data collection: towards parsing-based information extraction", 1st Int. Joint Conf. on Natural Language Processing (IJCNLP-04), (Asia Federation of Natural Language Processing), Hainan Island, China, Mar. (2004).

Suenaga A., Kiyatkin A. B., Hatakeyama M., Futatsugi N., Narumi T., Takada N., Ohno Y., Hoek J. B., Taiji M., Kholodenko B., and Konagaya A.: "Molecular dynamics simulation for investigation the structural impact of Shc phosphorylation", 1st Pacific-Rim Int. Conf. on Protein Science (PRICPS 2004), (Protein Science Society of Japan and others), Yokohama, Apr. (2004).

Fukuzaki A., Konishi F., Nagashima T., Ide K., Hatakeyama M., Yokoyama S., Kuramitsu S., and Konagaya A.: "Whole cell project of *Thermus thermophilus* HB8 toward atomic-resolution biology: Development of collaborative annotation system for *Thermus thermophilus* HB8 on OBIGrid", 1st Pacific-Rim Int. Conf. on Protein Science (PRICPS 2004), (Protein Science Society of Japan and others), Yokohama, Apr. (2004).

Schonbach C.: "Database strategies in immunology: turning data into information on the immune system", GBF Seminar, (GBF (German Research Centre for Biotechnology)), Braunschweig, Germany, Apr. (2004).

Schonbach C., Nagashima T., Matsuda H., Silva D., Petrovsky N., GER Group and GSL Members at RIKEN, and Konagaya A.: "From masking to inferring functional repeats in mouse", 9th Int. Human Genome Meet. (HGM2004), (HUGO), Berlin, Germany, Apr. (2004).

Demiya S., Schonbach C., Toyoda A., Nagashima T., Stahl U., Sakaki Y., Kuroki Y., and Fujiyama A.: "Screening of ancestral polymorphisms of immune response gene", 9th Int. Human Genome Meet. (HGM2004), (HUGO), Berlin, Germany, Apr. (2004).

Hosaka J., Kurochkin I. V., and Konagaya A.: "PBIE: a data preparation toolkit toward developing a parsing-based information extraction system", 4th Int. Conf. on Language Resources and Evaluation, (European Language Resources Association), Lisbon, Portugal, May (2004).

Schonbach C., Nagashima T., Kurochkin I. V., and Konagaya A.: "Computational identification and exploration of immune-related transcripts in mouse", 12th Int. Congr. of Immunology and 4th Ann. Conf. of FOCIS, Montreal, Canada, July (2004).

Kurochkin I. V., Schonbach C., and Konagaya A.: "The role of eukaryotic aminoacyl-tRNA synthetase complex in translation of α-helical regions of proteins", Biotechnology 2004, 12th Int. Biotechnology Symp. and Exh., (International Union of Pure and Applied Chemistry), Santiago, Chile, Oct. (2004).

Schonbach C.: "Variation and increased complexity of mammalian transcriptomes", International Workshop on: "Complexity and Modeling of Biological Networks", (GBF), Braunschweig, Germany, Oct. (2004).

（国内会議）

吉川澄美, 松村和美, 長嶋剛史, 保坂順子, 小長谷明彦: "薬物相互作用の固有表現分析", 言語処理学会第 10 回年次大会併設ワークショップ「固有表現と専門用語」, 東京, 3 月 (2004).

福崎昭伸, 小西史一, 長嶋剛史, 井手香, 畠山眞里子, 倉光成紀, 小長谷明彦: "Annotation object を用いた thermus thermophilus HB8 のアノテーションシステム", 第 28 回 JSAI SIG-MBI 分子生物情報研究会, 金沢, 3 月 (2004).

Schonbach C.: "In silico Biologie und deren Anwendung in der Medizin", Wissenschaftlicher Gespraechskreis, (German Embassy), 東京, 5 月 (2004).

長嶋剛史, Schonbach C., Kurochkin I. V., 小長谷明彦: "FREP: A system for inferring of functional repeats", CBI 学会 2004 年大会, 東京, 7 月 (2004).

福崎昭伸, 長嶋剛史, 井手香, 小西史一, 畠山眞里子, 横山茂之, 倉光成紀, 小長谷明彦: "Thermus thermophilus HB8 のゲノムワイドなアノテーション", CBI 学会 2004 年大会, 東京, 7 月 (2004).

Kurochkin I. V., 長嶋剛史, 小長谷明彦, Schonbach C.: "*In silico* discovery of peroxisomal proteome", CBI 学会 2004 年大会, 東京, 7 月 (2004).

柏原愛子, 石戸恵美, 吉良聡, 中川紀子, 井手香, 加納真, 長嶋剛史, 福崎昭伸, 小西史一, 畠山眞里子, 田代康介, 久原哲,

小長谷明彦, 横山茂之, 倉光成紀: “DNA マイクロアレイ解析”, 高度好熱菌丸ごと一匹プロジェクト第 3 回連携研究会, 播磨, 7–8 月 (2004).