C. 研究結果

1) HBV/G は単独では複製効率が悪く、他の genotype、特に HBV/Ae と共感染することで効率的な複製が可能となった。また、共感染により、肝繊維化が惹起されることもわかった。

2) HBV/A2 と B1_wild は肝障害性が低かったものの、それ以外のクローン群では星細胞が活性化しており、肝線維化が進行(stage F1-2)していた。さらに、線維化進行群では ground glass appearance がみられ HBV 抗原の細胞内への蓄積が見られ、強い肝障害が見られた。

D. 考察

HBV/G と HBV/Ae との共感染により効率的な複製が可能となり、肝組織進展に影響した。

また、HBV/C2 と B1_PC が強い肝障害を示し、臨床像と類似した結果となった。免疫能を欠くマウスのため、上記の結果は炎症を介さない HBV の直接的な肝障害性を示しているかもしれない。以上より、今回我々はヒト肝臓を用いた肝炎ウイルス病態モデルを新規に樹立した。これをさらに解析することで、肝癌などの肝病態進展に寄与する新たな病因や診断マーカーを探る手立てとなりうる。

E. 結論

キメラマウスが免疫抑制状態における臨床像と近い病態を示す新規の肝炎病態モデルマウスであることを確認した。

F. 健康危険情報

特になし。

G. 研究発表

1. 論文発表

Sugiyama M, Tanaka Y, Mizokami M. et al. Early Dynamics of Hepatitis B Virusin Chimeric Mice Carrying Human Hepatocytes Mono- or Coinfected with Genotype G. Hepatology. 45(4):929-937, 2007.

Sugiyama M, Tanaka Y, Mizokami M et al. Virological Properties of Distinct Hepatitis B Virus Genotypes in uPA/SCID Mouse with Human Hepatocyte. Submitted

2. その他の発表

Sugiyama M, Tanaka Y, et al. Differences of early dynamics and liver damage among hepatitis B virus genotypes in uPA/SCID mice with human hepatocytes. 58th Annual Meeting of the American Association for the Study of Liver Diseases. November 2-6 2007. Boston.

Tanaka Y, Sugiyama M, et al. Application of chimeric mice with human hepatocytes in hepatitis B virus research. APDW2007 Asian Pacific Digestive Week 2007. October 15-18 2007. Kobe.

Tanaka Y, Sugiyama M, et al. Hepatitis

B virus genotypes influence on viral replication and liver injury in vitro and in vivo. Presidential Prenary 1. 17th APASL Conference. March 27-30 2007. Kyoto.

Sugiyama M, Tanaka Y, et al. Viral Dynamics of Hepatitis B Virus in uPA/SCID Mice Carrying Human Hepatocytes Mono- or Coinfected with Genotype G. 17th APASL Conference. March 27-30 2007. Kyoto.

杉山真也，田中靖人，溝上雅史．　ワークショップ1：肝炎ウイルス研究の進歩．ヒト肝細胞置換キメラマウスでのHBV遺伝子型による複製効率の違い．第43回日本肝臓学会総会．平成19年5月31日-6月1日．東京．A24

H. 知的財産権の出願・登録状況
今回の研究内容については特になし。

厚生労働科学研究費補助金（肝炎等克服緊急対策研究事業）
（総括・分担）研究報告書

ジェノミクス技術を用いたウイルス性肝炎に対する新規診断・治療法の開発

分担課題：脂質代謝系解析を用いたウイルス性肝炎に対する医療開発研究

分担研究者　花田　賢太郎　国立感染症研究所細胞化学部長

C 型肝炎ウイルス（HCV）による難治性ウイルス肝炎の新規治療法の開発が強く求められている。HCV の産生及び病原性発現に関与する宿主脂質関連因子を探索するために、今回我々は、ウイルス産生及び病原性発現に重要な役割を果たしている HCV コアタンパク質に注目した。そこでコアタンパク質発現により変動する宿主脂質ラフト（界面活性剤不溶性画分）蛋白質の網羅的解析を培養肝細胞系を用い行った。各種変動蛋白質を同定し、その中からコアタンパク質発現細胞で発現が顕著に減少していた vimentin についてさらに解析を行った。その結果、vimentin 発現量の増加/抑制が HCV コアタンパク質量の抑制/増加を引き起こすことも明らかとなった。Vimentin はコアタンパク質と相互作用し、プロテアソーム依存的なコアタンパク質の分解に関与し宿主細胞内のコアタンパク質量に影響を与えていることが明らかとなった。培養細胞を用いた HCV 感染系の検討から、vimentin はコアタンパク質量の制御を介して HCV 産生にも影響を与えることが明らかとなった。このことから、Vimentin 及び vimentin-コアタンパク質相互作用部位が抗 HCV ターゲットとなる可能性が考えられた。

A．研究目的

難治性ウイルス肝炎の新規治療法の開発が強く求められている。HCV感染により肝脂肪化が特徴的に見られること、HCV複製と脂質代謝の関連が予想されるなど、HCVと脂質との関連が強く示唆されている。そこでHCV感染で発現が変動する脂質関連遺伝子を同定し、当該遺伝子が関わる脂質関連過程とHCV産生との関連を明らかにすることを目的とする。HCV産生に重要な脂質関連過程を見出せれば、それは新たな治療薬のターゲット候補になると期待される。

B．研究方法

HCV のコードする蛋白質の中でコアタンパク質はウイルス構造蛋白質としてウイルス産生に必須であるばかりでなく、コアタンパク質発現マウスでは脂肪蓄積・ガン化が見られることなどから病原性発現にも密接に関連している分子であることがわかってきている。そこで今回我々は HCV コアタンパク質に注目し、コアタンパク質発現により変動する宿主タンパク質のプロテオーム解析を培養肝細胞系を用い行った。培養細胞は、自ら樹立した Huh7 細胞由来の Uc321（コントロール用コアタンパク質非発現細胞）、Uc39-2 及び Uc39-6 細胞（コアタンパク質発現細胞）等を用いた。コアタンパク質の宿主細胞内主要局在部位である脂質ラフト画分（界面活性剤不溶性画分）をに焦点を当て比較プロテオーム解析を行った。具体的名方法は、以下のようである。各培養細胞を回収後 0.5% Triton-X100 を含む溶液で処理し、不溶性の画分を超遠心法等により生化学的に分離・精製を行った。分画した脂質ラフト画分中のタンパク質は二次元電気泳動法により分離し、染色後、コアタンパク質発現細胞・非発現細胞のタンパク質パターンを比較した。発現量に変動が見られたタンパク質スポットを採取し、トリプシン消化後、MALDI-TOF マススペクトロメトリーにより質量データを得、データベース検索によりタンパク質を同定した。

同定されたタンパク質分子の生化学的・細胞生物学的解析は以下のように行った。

各タンパク質の発現量の解析はイムノブロット法を用いた。mRNA 量の定量は RT-PCR 法を用い行った。

コアタンパク質との相互作用の解析は、免疫沈降法により行った。

コアタンパク質及び同定分子の細胞内過剰発現は pCAG あるいは pcDNA3.1 ベクターに当該遺伝子を組み込んだものをリポフェクション法にて行った。また、発現抑制は siRNA(dsRNA)法を用いて行った。

HCV感染培養細胞系は、Huh7細胞にHCV(JFH-1株)を感染させる系を用いた。感染能の測定には、HCV RNA 量の RT-PCR による定量、HCV コア蛋白質のイムノブロット法による検出および ELISA 法による定量、HCV コアタンパク質抗体による細胞免疫染色により行った。

（倫理面への配慮）

本研究ではヒト臨床材料・実験動物等を用いていない。そのために倫理面での問題はない。

C．研究結果

脂質ラフト画分（界面活性剤不溶性画分）の比較プロテオーム解析

Uc321 及び Uc39-6 細胞の脂質ラフト画分（界面活性剤不溶性画分）を用いた比較プロテオーム解析を行った。この画分には主に核タンパク質、細胞骨格系のタンパク質が濃縮される。解析の結果、Uc39-6 細胞（コアタンパク質発現細胞）において vimentin の顕著な減少、ケラチン 19 の増加を見いだした。また、複数の hnRNP 分子にも変動が見られた。

プロテアソーム依存的なコアタンパク質の分解における Vimentin の重要性

脂質ラフト画分において変動が顕著であった vimentin について更に解析を行った。

vimentin 量の減少は Uc39-2、Hep39 など複数のコア蛋白質発現細胞系でも再現された。コアタンパク質の過剰発現/抑制では vimentin の発現変動は見られず、逆に、vimentin の過剰発現/発現抑制で、コアタンパク質の発現量が有意に減少/上昇することが明らかとなった。これらの条件ではコアタンパク質の mRNA 量には変動が見られず、タンパク質レベルでの vimentin を介した制御が考えられた。宿主細胞内ではコアタンパク質はプロテアソーム依存的に速やかに分解されることが知られている。しかし、vimentin 発現抑制細胞ではプロテアソーム依存的なコアタンパク質の分解が強く阻害されていることが明らかとなった。また、vimentin とコアタンパク質が細胞内で相互作用することも明らかとなった。

vimentin 分子の HCV 産生への影響

以上の結果から、ウイルス感染培養細胞系を用い、ウイルス産生に対する vimentin 発現量の影響も検討した。vimentin 発現量を siRNA により抑制すると、ウイルス産生量が上昇し、ベクターにより vimentin を過剰発現した細胞ではウイルス産生量が低下していることが明らかとなった。

D．考察

本研究により複数のコアタンパク質発現細胞で vimentin 量の低下が認められた。細胞内 vimentin 量がコアタンパク質量に影響を与えることがわかったことから、我々が用いたコアタンパク質発現細胞株では vimentin 発現量が元々低かった為に、コアタンパク質量が高く維持されやすく分離に都合が良かった可能性が考えられる。コアタンパク質の安定発現株は非常に樹立しにくいことが知られている。一般に増殖時の培養細胞は vimentin の発現量が非常に高いので、このことと何らかの関連があるかもしれない。また、細胞増殖している肝がん部位では HCV 量が極めて低いことが知られているが、このこととも関係しているかもしれない。

vimentin 発現量がコアタンパク質量に影響を与えることから予想されるように、vimentin 発現量がウイルス産生にも影響を与えることがわかった。Vimenntin の発現抑制細胞では HCV 複製活性には影響が見られなかったことから、vimentin は恐らく主にコアタンパク質量への影響を介して感染能に影響を与えているものと考えられた。vimentin 量を抑制することにより HCV 産生能の高い培養細胞株を樹立できる可能性が考えられる一方で、逆に vimentin 量を上げる条件ではウイルス産生を抑制することができることを示した。以上より、vimentin、vimentin-コアタンパク質相互作用部位が抗 HCV 薬の標的となりうることが示唆された。

E．結論

プロテオミクスの手法を用い HCV コアタンパク質発現により変動する宿主脂質ラフト（界面活性剤不溶性画分）タンパク質の網羅的解析を行った。その結果、大きく変動する分子として vimentin を同定した。Vimentin はコアタンパク質と相互作用し、プロテアソーム依存的なコア蛋白質の分解に関与し宿主細胞内のコアタンパク質量に影響を与えることが明らかとなった。HCV 感染系を用いた検討から vimentin はコアタンパク質量の制御を介して HCV 産生に影響を与えることが明らかとなった。

F．健康危険情報

なし

G．研究発表

1．論文発表

なし

2．学会発表

1. Masayoshi Fukasawa, Yuko Nitahara-Kasahara, Fumiko Shinkai-Ouchi, Shigeko Sato, Tetsuro Suzuki, Kyoko Murakami, Takaji Wakita, Kentaro Hanada, Tatsuo Miyamura and Masahiro Nishijima, Cellular vimentin content affects the protein level of Hepatitis C virus core protein and the activity of Hepatitis C virus production in cultured cells. 14th International Symposium on Hepatitis C Virus and Related Viruses, Glasgow, UK, September, 2007
2. 小林翔、松田大介、深澤征義、西島正弘、花田賢太郎、司書毅、供田洋、ACC1 阻害剤による脂肪滴蓄積阻害活性、日本薬学会第１２８年会、横浜、2008年3月

G．知的所有権の取得状況

1．特許取得

なし

2．実用新案登録

なし

厚生労働科学研究費補助金（肝炎等克服緊急対策研究事業）
（~~総括~~・分担）研究報告書

ジェノミクス技術による新規薬剤リードの探索研究

（分担）研究者　菅　裕明　東京大学・先端科学技術研究センター・教授

研究要旨　菅研究室で独自に開発した特殊ペプチド合成法（RaPI Dシステム）を用いて、本研究班で新規に同定された標的に対する阻害剤を探索する。本年度は、その下準備として、ライブラリー合成に向けた技術基盤の確立に挑んだ。

A．研究目的

本研究班では、ウィルス性肝炎および肝癌における病態を網羅的なジェノミクス手法を用いて解析し、新規標的もしくは疾患関連因子の同定を目標としている。本研究分担班は、それらの新規標的や疾患関連因子に対し、薬物リードを迅速に発見を目指し、関連技術を開発し、実施することを目的としている。

B．研究方法

当研究室で独自に開発したRaPID（Random Peptide Integrated Discovery）システムを駆使し、翻訳系を用いて特殊ペプチドを合成、薬剤探索にあてる。本技術の特筆すべき点は、フレキシザイム（tRNAアミノアシル化RNA触媒）を用いて、普遍遺伝暗号表を初期化し、通常アミノ酸を異常アミノ酸に対応させた改変遺伝暗号表を作成し、それに沿った形でmRNAを翻訳することで特殊アミノ酸を合成する手法である。

（倫理面への配慮）

これまでのところ、本研究はヴィトロを中心としており、倫理面への配慮を特にない。

C．研究結果

本年度は、特殊ペプチドライブラリー合成のための基盤技術の確立を目指した。特筆すべき成果として、それぞれ異なる特色持った環状化特殊ペプチドの合成法を３手法開発した。そのうちの１つの手法に関しては、GPCRファミリーのウロテンシンレセプターに結合するウロテンシンIIの特殊ペプチド化に応用し、活性を保持しながらプロテアーゼ耐性を植え付けることに成功した。この成果は、特殊ペプチドが薬剤開発に適していることを強く示唆しており、今後の薬剤探索を進める上で大きな技術基盤となる。

D．考察

本技術の最大の特徴は、ライブラリー合成の迅速性である。そのパイロットして３２０種類の特殊ペプチド合成を既に検討し、準備期間も含め１週間という短期間で達成できた。平成２０年度は、より高い多様性をもったライブラリー構築を行い、ウィルス性肝炎および肝癌の新規標的もしくは疾患関連因子に対する薬剤開発に移行したい。

E．結論

目標の技術開発は達成した。

F．健康危険情報

G．研究発表

1.　論文発表

・A. Ohta, H. Murakami, E. Higashimura, H. Suga“Synthesis of polyester by means of genetic code reprogramming” *Chemistry & Biology*, 14, 1315-1322 (2007).

・T. Kawakami, H. Murakami, H. Suga “Messsnger RNA-directed incorporation of multiple N-methyl-amino acids into linear and cyclic peptides”*Chemistry & Biology*, 15, 32-42 (2008).

・Y. Goto, A. Ohta, Y. Sako, Y. Yamagishi, H. Murakami, H. Suga “Reprogramming the initiation event in translation for the synthesis of physiologically stable cyclic peptides”*ACS Chemical Biology*, 3, 120-129 (2008).

・Y. Sako, Y. Goto, H. Murakami, H. Suga “Ribosomal synthesis of peptidase-resistant peptides closed by a non-reducible inter-sidechain bond ”*ACS Chemical Biology*, 3, 120-129 (2008).

2. 学会発表

・ 2007 年 5 月 14 日 : Protein Engineering Summit, Boston, USA

・2007年5月24日：第７回日本蛋白質科学会年会、仙台

・ 2007 年 6 月 22 日 : EMBO Workshop, Chemistry and Biochemistry of Enzyme Catalysis by Biological Systems, EMBL-Hamburg, Germany

・2007年6月25日：EPFEL Chemical Biology Symposium, Lausanne, Switzerland

・ 2007 年 7 月 6 日 : 50th Annual Meeting, Systems and Chemical Biology, Canadian Society of Biochemistry, Molecular & Cellular Biology, Montreal, Canada.

H．知的財産権の出願・登録状況
（予定を含む。）
本プロジェクトに関しての知的財産権は本年度はなし。

厚生労働科学研究費補助金（肝炎等克服緊急対策研究事業）
分担研究報告書

統計的因果推定法に基づくジェノミクス解析法の研究

（分担）研究者　堀本　勝久　産業技術総合研究所・研究チーム長

研究要旨
グラフィカルモデルに基づく統計的なネットワーク解析手法を肝がん細胞において計測されたデータについて適用し、肝がん特異的関連ネットワーク及び進展過程におけるネットワーク変化の同定することで、疾患機序の解明のための実験支援を行う。

A．研究目的

統計的手法に基づき、計測データからネットワーク推定及び評価の方法をマイクロアレイデータなどの計測データに適用し、肝がん特異的遺伝子ネットワークの構築する。

B．研究方法

グラフィカルモデルに基づくネットワーク推定法を肝がんについて計測されたデータに適用し、遺伝子間の関連性をグラフ表現する。

（倫理面への配慮）

当機関においては、理論研究のみを実施するため、動物実験は該当無し。

C．研究結果

遺伝子発現計測データと文献情報に基づく既知ネットワーク構造との整合性を見積もる方法をグラフィカルモデルに基づいて開発した。整合性評価の性能評価のため、大腸菌ゲノムについて嫌気性条件下で計測された遺伝子発現データと３０の既知ネットワークに適用し嫌気性下で活性化されているネットワークを選択的に同定することに成功した。

D．考察

特異的な条件下で計測されたデータを用いて活性化ネットワークを推定する手法を、疾患細胞について計測されたデータに適用することで、疾患特異的な活性化ネットワークの同定が可能であることを示唆する。計測データに関して整合性を見積もる既知ネットワークデータの整備を広く行い、漏れのない評価を実施することが肝要である。

E．結論

未知の疾患特異的遺伝子間ネットワークを統計的手法に基づき推定することで、実験研究との連携により疾患機序の解明を加速する。

F．健康危険情報

G．研究発表

1.　論文発表

Aburatani, S., et al., *EURASIP J. Bioinfo. Systems Biol.*, 47214, 2007.

Yoshida, H., et al., *BioSystems*, 90, 486–495, 2007.

2.　学会発表

Sato, T., et al. *Algebraic Biology* July, Linz, Austria, 2007

Yoshida, T.,et al. 10th *CASC,* Sug., Bonn, Germany, 2007

H．知的財産権の出願・登録状況

（予定を含む。）

1. 特許取得

なし

2. 実用新案登録

なし

3.その他

なし

III. 研究成果の刊行に関する一覧表

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書　籍　名	出版社名	出版地	出版年	ページ
Sato,T,（堀本）	Inference of Protein-Protein Interactions by Using Co-evolutionary Information	Anai,H, Horimoto, K,Kutsia,T	Algebraic Biology 2007 (Lecture Notes in Computer Science 4545)	Springer	Heidelberg	2007	322-333
Yoshida, H,（堀本）	Exact parameter determination for Parkinson's disease diagnosis with PET using an algebraic approach	Anai,H, Horimoto, K,Kutsia,T	Algebraic Biology 2007 (Lecture Notes in Computer Science 4545)	Springer	Heidelberg	2007	110-124
Yoshida, H,（堀本）	An Algebraic-Numeric Algorithm for the Model Selection in Kinetic Networks	Ganza,GV, Mayer,EW, Vorozhtsov, E	CASC 2007 (Lecture Notes in Computer Science 4770)	Springer	Heidelberg	2007	433-447
Hayashida,（堀本）	Integer Programming-based Approach to Allocation of Reporter Genes for Cell Array Analysis	Zhang,X-S, Chen,L,Wu,L-Y, Wang,Y	OSB 2007 (Lecture Notes in Operation Research 7)	World Publishing Corporation	Beijing	2007	288-301

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Minagawa H,（金子）	Comparative proteomic and transcriptomic profiling of the human hepatocellular carcinoma.	Biochem Biophys Res Commun	366(1)	186-192	2008

Matsuzawa N, (金子)	Lipid-induced oxidative stress causes steatohepatitis in mice fed an atherogenic diet.	Hepatology	46(5)	1392-1403	2007
Oishi N, (金子)	Hepatitis B virus X protein overcomes oncogenic RAS-induced senescence in human immortalized cells.	Cancer Sci	98(10)	1540-1548	2007
Takamura T, (金子)	Gene expression profiles in peripheral blood mononuclear cells reflect the pathophysiology of type 2 diabetes.	Biochem Biophys Res Commun	361(2)	379-384	2007
Komura T, (金子)	Impact of diabetes on recurrence of hepatocellular carcinoma	Am J Gastroenterol	102(9)	1939-1946	2007
Hiraga N, (金子)	Infection of human hepatocyte chimeric mouse with genetically engineered hepatitis C virus and its susceptibility to interferon.	FEBS Lett	581(10)	1983-1987	2007
Tateno M, (金子)	Expression profiling of peripheral-blood mononuclear cells from patients with chronic hepatitis C undergoing interferon therapy.	J Infect Dis	195(2)	255-267	2007
Jinushi M, (竹原)	Natural killer cell and hepatic cell interaction via NKG2A leads to dendritic cell-mediated induction of CD4+ CD25+ T cells with PD-1-dependent regulatory activities.	Immunology	120	73-82	2007

Sakamori R, (竹原)	Signal transducer and activator of transcription 3 signaling within hepatocytes attenuates systemic inflammatory response and lethality in septic mice.	Hepatology	46	1564-1573	2007
Hiramatsu N, (竹原)	Early decline of hemoglobin can predict progression of hemolytic anemia during pegylated interferon and ribavirin combination therapy in patients with chronic hepatitis C.	Hepatol Res	38	52-59	2008
Matsumoto Y, (坪内)	Inhibition of tumor-stromal interaction through HGF/Met signaling by valproic acid.	Biochem Biophys Res Commun	366	110-169	2008
Uto H, (坪内)	Alanine aminotransferase flare-up in hepatitis C virus carriers with persistently normal alanine aminotransferase levels in a hyperendemic area of Japan.	J Gastroenterol	42	673-680	2007
Kanmura S, (坪内)	Early diagnostic potential for hepatocellular carcinoma using the SELDI ProteinChip system.	Hepatology	45	948-956	2007
Abe H, (坪内)	Transgenic expression of osteoactivin in the liver attenuates hepatic fibrosis in rats.	Biochem Biophys Res Commun	356	610-615	2007

宇都浩文、(坪内)	C 型肝炎治療 up to date;C 型慢性肝炎-ALT 持続正常者の治療戦略-.	Mebio	25	82-87	2007
井戸章雄、(坪内)	特集肝不全の各種病態と新しい治療の視点 2;肝再生誘導のスイッチとそれを妨げる因子.	Surgery Frontier	14	133-137	2007
Tasaka M, (前川)	HCV nonstructural proteins responsible for suppression of RIG-I/Cardif -induced interferon response.	J Gen Virol.	88	3323-33	2007
Sakamoto N, (前川)	Inhibition of hepatitis C virus infection and expression in vitro and in vivo by recombinant adenovirus expressing short hairpin RNA.	J Gastroenterol Hepatol.	-	-	2007
Amemiya F, (前川)	Targeting lipid metabolism in the treatment of hepatitis C.	J Infect Dis	197(3)	361-70	2008
Jin H, (前川)	Griseofulvin, an oral antifungal agent, suppresses HCV replication in vitro.	Hepatol Res	In Press		2008
Sugiyama M, (田中)	Early Dynamics of Hepatitis B Virus in Chimeric Mice Carrying Human Hepatocytes Mono- or Coinfected with Genotype G.	Hepatology	45(4)	929-937	2007
A. Ohta (菅)	Synthesis of polyester by means of genetic code reprogramming	Chemistry & Biology	14	1315-1322	2007
T. Kawakami (菅)	Messsnger RNA-directed incorporation of multiple N-methyl-amino acids into linear and cyclic peptides	Chemistry & Biology	15	32-42	2008

Y. Goto (菅)	Reprogramming the initiation event in translation for the synthesis of physiologically stable cyclic peptides	ACS Chemical Biology	3	120-129	2008
Y. Sako (菅)	Ribosomal synthesis of peptidase-resistant peptides closed by a non-reducible inter-sidechain bond	ACS Chemical Biology	3	In press	2008
Aburatani, S., (堀本、金子)	Gene systems network inferred from expression profiles in hepatocellular carcinogenesis by graphical Gaussian model	EURASIP J Bioinfo Systems Biol	47214	1-11	2007
Yoshida, H., (堀本)	Derivation of rigorous conditions for high cell–type diversity by algebraic approach	BioSystems	90	486–495	2007

Ⅳ．研究成果の刊行物・別刷

Hirokazu Anai · Katsuhisa Horimoto
Temur Kutsia (Eds.)

Algebraic Biology

Second International Conference, AB 2007
Castle of Hagenberg, Austria, July 2-4, 2007
Proceedings

Springer

Inference of Protein-Protein Interactions by Using Co-evolutionary Information

Tetsuya Sato[1], Yoshihiro Yamanishi[2], Katsuhisa Horimoto[3], Minoru Kanehisa[2], and Hiroyuki Toh[1]

[1] Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka 812-8582, Japan
sato@bioreg.kyushu-u.ac.jp, toh@bioreg.kyushu-u.ac.jp
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
yoshi@kuicr.kyoto-u.ac.jp, kanehisa@kuicr.kyoto-u.ac.jp
[3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan
k.horimoto@aist.go.jp

Abstract. The mirror tree is a method to predict protein-protein interaction by evaluating the similarity between distance matrices of proteins. It is known, however, that predictions by the mirror tree method include many false positives. We suspected that the information about the evolutionary relationship of source organisms may be the cause of the false positives, because the information is shared by the distance matrices. Therefore, we excluded the information from the distance matrices and evaluated the similarity of the residuals as the intensity of co-evolution. We developed two different methods with a projection operation and partial correlation coefficient. The number of false positives were drastically reduced by our methods.

Keywords: protein-protein, co-evolution, projection operation, partial correlation coefficient.

1 Introduction

Information about protein–protein interactions in living cells provides deep insight into the biological functions of proteins at the cellular level. The development of large-scale experimental analyses, such as the yeast 2-hybrid system [7,21] and pull-down method [3,6], has facilitated understanding the protein–protein interaction network in cells. However, such experimental approaches have problems in coverage and accuracy [20,22]. Following the trend, the prediction of protein–protein interactions has become one of the major issues in bioinformatics. The predicted protein–protein interactions can provide complementary or supporting evidence to the large-scale experimental studies on protein–protein interactions although computational analyses also have the same drawbacks as experimental studies, that is, low coverage and low accuracy.

H. Anai, K. Horimoto, and T. Kutsia (Eds.): AB 2007, LNCS 4545, pp. 322–333, 2007.

Various computational methods to predict protein-protein interactions have been developed until today. Co-evolutionary behavior between interacting proteins provides useful information for the prediction of protein-protein interaction. The mirror tree method [15] and the *in silico* 2-hybrid system method [14] are two representative methods to predict protein-protein interaction with co-evolutionary information. In this paper, we explain our studies [18,19] aiming at improvement of the mirror tree method. The mirror tree method was developed by Pazos and Valencia [15], although there are several preceding works, such as Goh *et al.* [5]. The mirror tree method predicts protein–protein interactions under the assumption that the interacting proteins show similarity in molecular phylogenetic tree because of the co-evolution through the interaction. To avoid the difficulty to evaluate the similarity between a pair of phylogenetic trees, however, the mirror tree method compares a pair of distance matrices. Consider two proteins, proteins A and B. The orthologous amino acid sequences of protein A are collected from n species. The n sequences of protein A are aligned and the distance matrix, D_{A}, is calculated. The size of D_{A} is $n \times n$, and each row or column of the matrix corresponds to a species under consideration. An element of the matrix, $D_{\mathrm{A}}(i,j)$, represents the genetic distance between species i and j, which is calculated by comparing the amino acid sequences of protein A between the two species. A distance matrix is symmetric, and only the upper or lower half of the matrix includes sufficient information for tree construction. Likewise, the orthologous amino acid sequences of protein B are collected from the same n species, and the distance matrix, D_{B}, is calculated. The intensity of co-evolution between proteins A and B is evaluated as Pearson's correlation coefficient, $\rho_{\mathrm{AB}}^{\mathrm{MIRROR}}$, between the distance matrices D_{A} and D_{B}, which is calculated as follows:

$$\rho_{\mathrm{AB}}^{\mathrm{MIRROR}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (D_{\mathrm{A}}(i,j) - \mathrm{Ave}(D_{\mathrm{A}}))(D_{\mathrm{B}}(i,j) - \mathrm{Ave}(D_{\mathrm{B}}))}{\sqrt{\mathrm{Var}(D_{\mathrm{A}})\mathrm{Var}(D_{\mathrm{B}})}}, \quad (1)$$

where Ave and Var represent the average and the variance of the upper (or lower) half elements of a distance matrix. High correlation between the distance matrices indicates the resemblance of the corresponding phylogenetic trees. Therefore, a pair of proteins are predicted to interact with each other, when the distance matrices of the proteins show high correlation. Because of the simplicity, modification and improvement have been introduced into the mirror tree method by several groups [4,9,16]. On the other hand, it has been recognized that the mirror tree predictions include many false positives. That is, even protein pairs that are known not to interact often show high correlation coefficients. Then, such pairs are predicted to interact in error. The abundance of false positives in the mirror tree prediction reduces the reliability of the method in actual applications. We suspected that the cause of the false positives is the information about the evolutionary relationship among the source organisms of the collected orthologous sequences. The distance matrices of orthologous proteins from the same set of n source organisms are compared in the mirror tree method. Therefore, all of the

distance matrices of the proteins are considered to include the same information about the evolutionary relationships among the same n sources. The information shared by the distance matrices would generate high correlation even between the matrices of non-interacting proteins. If our hypothesis is correct, the number of false positives in the predictions could be reduced by excluding such information from the distance matrices. We developed two different methods to exclude the information from the distance matrices for the prediction of protein-protein interaction. One of them uses a projection operator, whereas the other is based on multiple regression. The two methods were applied to physically contacting proteins, to evaluate their performances. Then, it was found that our methods drastically reduced the number of false positives in the predicted protein-protein interactions as expected.

2 Material and Method

2.1 Data Preparation

13 pairs of *Escherichia coli* proteins that are physically in contact were selected from the Database of Interacting Proteins (DIP) [17]. The pairs are listed in the legend for Table 1. Each pair was selected so that neither of the interacting proteins participated in the remaining 12 pairs of interacting proteins. Then, (putative) orthologues corresponding to the 26 proteins were collected from 40 different bacterial species, according to the KEGG KO database [10]. We assumed that a pair of proteins, which are orthologous to the interacting proteins of *E. coli*, are also physically in contact.

2.2 Multiple Sequence Alignment and Distance Matrix

A multiple alignment of each set of orthologous amino acid sequences was made with the alignment software MAFFT [11]. A distance matrix for the orthologous sequences was calculated from the multiple alignment. A genetic distance between every pair of aligned sequences was calculated as a maximum likelihood estimate using the PROTDIST in the PHYLIP package [2]. JTT model [8] was used as a model for the amino acid substitution for the estimation.

2.3 Transformation from Distance Matrix to Phylogenetic Vector [19]

The distance matrix was transformed into a vector. The upper or lower half of the non-diagonal elements of the distance matrix was arranged as a one-dimensional array of the numerical values in a certain order. All of the matrices were transformed into vectors with the same arrangement of the elements. When the matrix has a size of $n \times n$ the dimension of the vector is $n(n-1)/2$. The vector is hereafter referred to as a 'phylogenetic vector'. The dimension of the phylogenetic vector is 820, because n is 41. Consider a pair of phylogenetic vectors, which are transformed from distance matrices D_i and D_j. The subscripts

i and j indicate different sets of orthologues, that is, different proteins. Then, the elements of each vector are normalized with the average and the standard deviation of the elements as follows:

$$\left|\nu_i^{\#}\right\rangle = \frac{|\nu_i\rangle - |\mu\rangle}{\sqrt{\mathrm{Var}(\nu_i)}}, \tag{2}$$

where $|\mu\rangle$ is a vector with the same dimension as $|\nu_i\rangle$. All the elements of $|\mu\rangle$ are constant, and are equal to the arithmetic average over the elements of $|\nu_i\rangle$. $\mathrm{Var}(\nu_i)$ is the variance over all the elements of $|\nu_i\rangle$. The superscript # in $\left|v_i^{\#}\right\rangle$ indicates that the vector is normalized. Then, the inner product between a pair of normalized vectors is the Pearson's correlation coefficient used for the mirror tree method, which is defined by formula (1). Hereafter, the correlation coefficient by the mirror tree method is denoted as $\rho_{ij}^{\mathrm{MIRROR}}$.

$$\rho_{ij}^{\mathrm{MIRROR}} = \langle \nu_i^{\#} | \nu_j^{\#} \rangle. \tag{3}$$

2.4 First Method with Projection Operator [19]

Consider an $n(n-1)/2$-dimensional unit vector $|u\rangle$, which represents the evolutionary relationship of the source species under consideration. Given such a vector, following projection operator P can be defined:

$$P = I - |u\rangle \langle u| . \tag{4}$$

The projection operator is a matrix with the size of $n(n-1)/2 \times n(n-1)/2$. The method to obtain $|u\rangle$ is explained below. I represents an identity matrix with the size of $n(n-1)/2 \times n(n-1)/2$. By applying the projection operator (4) to a phylogenetic vector, say, $|\nu_i\rangle$, the component within $|\nu_i\rangle$, which is orthogonal to $|u\rangle$, is generated:

$$|\varepsilon_i\rangle = P|\nu_i\rangle = |\nu_i\rangle - |u\rangle \langle u|\nu_i\rangle . \tag{5}$$

$|\varepsilon_i\rangle$ is a residual vector obtained by excluding the information about the evolutionary relationship from the phylogenetic vector . The same projection operator was applied to all of the phylogenetic vectors under consideration. Each of the residual vectors was then normalized with the average and the standard deviation of the elements. The inner product between the two residual vectors $\left|\varepsilon_i^{\#}\right\rangle$ and $\left|\varepsilon_j^{\#}\right\rangle$ represents the Pearson's correlation coefficient between the residual vectors:

$$\rho_{ij}^{\mathrm{PROJECTION}} = \langle \varepsilon_i^{\#} | \varepsilon_j^{\#} \rangle \tag{6}$$

was used as a new measure to evaluate the intensity of co-evolution between proteins i and j.

In order to obtain the unit vector representing the phylogenetic relationship of the source organisms, three different methods were considered. In the first method, 16S rRNA was used for the calculation. Basically, at least one copy of the 16S rRNA gene is encoded by each genome. Therefore, the distance matrix or the phylogenetic vector of the 16S rRNAs is considered to represent the evolutionary relationship among the source organisms. The nucleotide sequences of rRNA were collected from the same sources as the proteins under consideration according to the KEGG GENES database [10] and the Ribosomal Database Project-II Release 9 [1]. The nucleotide sequences of the 16S rRNA were aligned, and the distance between every pair of the aligned nucleotide sequences was calculated by using the F84 model [12] with the DNADIST in the PHYLIP package [2]. The distance matrix was then transformed into a phylogenetic vector $|\nu_{16S}\rangle$. Then, a unit vector $|u_{16S}\rangle$ was obtained as $|\nu_{16S}\rangle/\|\nu_{16S}\|$.

In the second method, all of the phylogenetic vectors of proteins under consideration were normalized so that the size of the elements in each protein was '1' at first. Then, they were averaged as

$$|\nu_{\mathrm{AVE}}\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{|\nu_i\rangle}{\|\nu_i\|}, \tag{7}$$

where m is the number of proteins. So, m was 26 here. The second unit vector $|u_{\mathrm{AVE}}\rangle$, was obtained as $|\nu_{\mathrm{AVE}}\rangle/\|\nu_{\mathrm{AVE}}\|$.

In the third method, the phylogenetic vectors were used again. Let X be a matrix of $n(n-1)/2 \times m$ in which the i-th column corresponds to a normalized phylogenetic vector of protein i. Then, a correlation coefficient matrix Y of $m \times m$ was calculated as $X^T X$. The superscript T indicates the transpose of a matrix. The principal component analysis for the data corresponding to X is equivalent to solving the eigenvalue problem of Y. Then, $|\nu_{\mathrm{PC1}}\rangle$ was obtained as $|\nu_{\mathrm{PC1}}\rangle = X|z_1\rangle$, where $|z_1\rangle$ is a vector corresponding to the first principal component axis. Then, $|\nu_{\mathrm{PC1}}\rangle/\|\nu_{\mathrm{PC1}}\|$ generated the third unit vector, $|u_{\mathrm{PC1}}\rangle$.

The Pearson's correlation coefficients between the residual vectors for a pair of proteins i and j, which were generated by the projection operations constructed with $|u_{16S}\rangle$, $|u_{\mathrm{AVE}}\rangle$ and $|u_{\mathrm{PC1}}\rangle$, were represented by ρ_{ij}^{16S}, ρ_{ij}^{AVE} and ρ_{ij}^{PC1}. The type of correlation coefficient is collectively represented by ρ^* without the subscripts, i and j where the superscript indicates the type of correlation coefficient.

2.5 Second Method with Multiple Regression [18]

Suppose that m proteins are given and we want to predict interacting pairs from them. Consider multiple regressions of $|\nu_i\rangle$ and $|\nu_j\rangle$ with $(m-2)$ phylogenetic vectors:

$$|\nu_i\rangle = \alpha_0 + \sum_{k\neq i,j}^{m}\alpha_k|\nu_k\rangle + |\delta_i\rangle, \tag{8}$$

$$|\nu_j\rangle = \beta_0 + \sum_{l\neq i,j}^{m}\beta_l|\nu_l\rangle + |\delta_j\rangle, \tag{9}$$