1G3-2

遺伝子データからの相関する概念抽出と関係づけオントロジーの作成

Ontology construction focused on relationships between correlated concepts in gene databases

村上 勝彦[*1]
Katsuhiko Murakami

今西 規[*2]
Tadashi Imanishi

[*1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)

[*2] 東海大学
Tokai University

Construction of ontology is a fundamental issue in semantic web. In molecular biology, when we integrate genetic data from different sources, there are many same, similar, or related concepts that were independently created and annotated. To integrate the similar but different concepts, we extracted them and introduced some properties that can describe the relationships between similar concepts precisely. The enriched integrated ontology will contribute to more sophisticated information processing or data mining.

1. はじめに

分子生物学でもセマンティック Web 技術を用いてデータの統合や統合的解析に役立てようという研究活動が活発である[山口 11]. 分子生物学で重要な遺伝子等の分子データは, その分子 ID と分子が何であるかの記述, どういう働き(医学生物学的または生化学的役割・機能)があるかを記述するためのデータ ID とタームが, データベース毎に作成されて付与されている. これらに対する統合的に情報処理解析をしたい時, ことなる ID が示す分子や機能の意味が同じなのか, 近いのか, またはそれ以外のどのような関係なのかをなるべく正確に把握し, 記述するような形でのデータ統合が必要になる. なお以下では複雑な表記を避けるため, 特別な場合を除き, タンパク質や転写産物を「遺伝子」と表記している.

遺伝子等の分子においては, INSDB のアクセッション番号などの ID によって, 比較的精密に区別が可能である. しかし, 各遺伝子の特徴を示す用語(例えば「酵素活性」といった機能を示す何か)においては, さまざまな粒度で異なる表現が存在し利用されている. 遺伝子機能を示す概念で有名な例は, Gene Ontology (GO, [GO 03]) である. GO では 3 種の概念, すなわち機能(molecular function), 生物学的な反応過程(biological process), および細胞構成要素(cellular component)についてのオントロジーが提供されている. GO を使って記述可能な概念は生命現象の一部に過ぎないので, 各データベースや解析ツールは独自の概念を自由に各遺伝子に付与している.

バイオデータベースでは一般に, 各遺伝子に GO などの概念が付与されているが, 中にはテキスト注釈されただけの「浮いている概念」も存在する. これらはこのままでは他の概念とどのような関係にあるか不明である. そのため機械的な利用は簡単ではない. 統合的な情報処理をするには, 現状の「人の判断(読解)を前提としたテキスト記述や ID の羅列」から, それらを意味的につなげなければならない. このために我々は, 適切な Property を用いて, 異なるデータベースで生成された概念関係を把握して明示することにより, データの高度利用を目指している. 従って, 2つのオントロジー間で同じ概念(クラス)を自動的にマッチさせるいわゆるオントロジーマッピング[Euzenat 13]とは目標が異なる. ここでの目的は, 似ている概念がどう似ているかを分類, 考察して, (異なる概念として)適切な関係を付与することである. オントロジー構築過程の途中に人の目視確認や判断を加えて, 最終的には間違いのない(または信頼度が付与された)統一的オントロジーを構築し, 将来的にそれをマイニングなどのアプリケーションに利用することを目指している.

連絡先:村上勝彦, 産業技術総合研究所 創薬分子プロファイリング研究センター, 135-0064 東京都江東区青海二丁目4番7号, k-murakami@aist.go.jp, aaaccc.k@gmail.com

2. データと方法

2.1 ヒト遺伝子データ

ヒト遺伝子データセットとして H-InvDB [Imanishi 04, Takeda 12] (Release 8.3)を利用した. このデータは, 各遺伝子を主キーとして、さまざまな(外部データベースに由来する)概念を付与している. 異なる概念でも意味は一部オーバーラップしている. 今回使った遺伝子の特徴を示す概念の種類(括弧内は概念由来のデータベースやソフト)は, gene family (H-InvDB), Gene Ontology (GO), 機能ドメイン (InterPro), 代謝経路 (KEGG pathway), タンパク質間相互作用(H-InvDB), 立体構造モチーフ (SCOP), 疾患 (OMIM), 組織特異的な遺伝子発現をする組織 (H-InvDB), および細胞内局在(H-InvDB)の 9 種類である. 各概念の値としては, 例えば OMIM であれば「II 型糖尿病」などの具体的疾患名が示され, その遺伝子はその疾患に関わる遺伝子であることが示されている. H-InvDB には, 6 段階の遺伝子信頼性を示すカテゴリがあるが, 今回は H-InvDB カテゴリ I の遺伝子, すなわちタンパク質の存在が実験で確認されている 16,138 個の遺伝子を用いた. 各遺伝子には, 上記の概念が 0 個以上付与されている. ほとんどすべての概念は排他的な関係ではない. 表1に概念の数を示す.

表1 遺伝子に付与された概念の種類と名称のユニーク数

番号	概念種類	ユニークな名称数	付与された遺伝子数
1	組織特異的遺伝子発現	10	1,238
2	細胞内局在	11	14,108
3	代謝経路(KEGG)	168	847
4	機能等(GO)	1,696	10,046
5	疾患(OMIM)	2,068	1,752
6	蛋白立体構造(SCOP)	2,232	10,620
7	Gene Family	2,863	9,741
8	機能ドメイン(InterPro)	6,615	12,764
9	蛋白相互作用	9,945	7,528

2.2 相関する概念の網羅的検出と概念間の関係性を示すオントロジーの作成

概念間で関連するものをリストアップするために，相関の強さを調べた．各概念ペアを毎にその概念を付与された遺伝子数に基づき，図1のような分割表を作成して，概念の独立検定をFisher exact test（両側）で P 値を計算する．Bonferroni の多重検定で5％以下のペアを有意に相関があると設定した．その結果，99,747 ペアに正の相関がみられ，526 ペアに負の相関がみられた．今回は正の相関について報告するが，強い正の相関を示すペアの多くは，同じ概念に関するものが多かった．

1. SC: COOP Family g.44.1.1 (RING finger domain)
2. IRP: IPR001841 (Zinc finger, RING-type)

		SC=T	F
IRP	T	157	71
	F	35	15,875

P-value 2.1e-284

図1　SCOP:g.44.1.1 と IPR:001841 の概念間の分割表と P 値．分割表の左上数値は IPR ドメインと SCOP ドメインの両方が付与された遺伝子数，右下はどちらも付与されてない遺伝子数に対応する．T/F は概念付与の真偽を示す．

強い相関を示すペアの1つの例は，"SCOP g.44.1.1 (RING finger)" と "IPR001841 (Zinc finger, RING-type)"（図1）である．これらは文字列からもわかるようにタンパク質配列のパターンが既知の特徴をもっていることを示すものであり，同じ概念である．これらはデータ統合という観点から同じ概念クラスとして扱いたい．そこで，2つの方法が考えられる．1つ目は，相関する2つの概念を skos:closeMatch [W3C 09]でつなぐことである．これは一部の情報検索アプリケーションで交換して使用できることを示す．これは強い相関で，かつ文字列も似ている場合には適当である．しかし，付与されている概念には予測によるものが含まれるため，すべて含めて同じように表現してしまうと，弱い証拠しかない場合でも強い証拠の場合と同様に遺伝子に情報付与してしまうという危険性が高くなる．そこで2つ目の方法として，統合的なスーパークラス”hinvo:RING-type zinc finger”を作成して，そのサブクラスとして SCOP（立体構造モチーフ）や IPR（機能ドメイン）を位置づける方法が考えられる．加えて「クラス間の距離」（例えば Jaccard 係数）を遺伝子カウントによって付与する方法が考えられる．ここで"hinvo:"とは，本研究独自のオントロジーであることを示す．これによって，概念間の距離を付与すれば，ユーザーは2つの概念の同等性を自由に扱うことが出来る．クラス間の距離が連続値であることは，ユーザー側で閾値を調整できるので，さらに高度な解析や予測に使えるというメリットがある．

相関のある概念のうち意味が同一でないものには，一見関係なさそうで実は深い関連があるタイプがあった．例として，dephosphorylation（脱リン酸化）と1型糖尿病の例を図2に示す．この関連は少数の文献にしか掲載されていないため，一般の研究者にとってこのタイプの関連性を提示することには意義があると思われる．このタイプは，hinvo:correlatedWith というプロパティで関連づけた．

相関を示すもので別の興味深いパターンは，ある概念を付与された遺伝子の全て（ここでは例外のない場合を考える）が，別の概念を付与されているケースである．所謂「相関ルール」で，

1. GO: GO:0006470, protein dephosphorylation
2. KG: hsa04940 Type I diabetes mellitus

	KG=T	F
GO=T	47	37
F	5	16,049

P-value 1.1e-108

図2　GO:0006470 と KEGG, hsa04940 についての分割表と P 値．分割表の左上数値は GO タームと KEGG パスウェイの両方が付与された遺伝子数，右下はどちらも付与されてない遺伝子数に対応する．T/F は概念付与の真偽を示す．

A ならば B という関係である．例えば「HIP00006060 (RAD51B, DNA 修復に関わる蛋白質) と相互作用する蛋白質は，全て GO:0006334:nucleosome assembly (Cellular Process の GO)が付与されている」というルールが抽出された（P-value 1.9e-77）．ある概念が高い確率で別の概念を意味するということから hinvo:imply としたいところであるが，これら2つの概念の関連は示されていたものの，詳細なメカニズムは現在生化学的に研究されているところである[North 13]．「予測された関連性がある」という以上に強くいうことは危険かもしれないが，関連付け程度であれば妥当と解釈できる．方向性があることを考慮し、このタイプの関係性は，hinvo:mayImply として関連づけた．

3. 議論

本研究ではヒト遺伝子の統合データベース H-InvDB に付与されたもともと由来が異なる概念について相関のあるものを具体的に抽出し，それらの関連性を示すオントロジー(property)を提案した．

今後，異なるデータベースを統合していく場合でも同様のアプローチが可能である．概念の類似度データを付与する方法は，相関ルール発見などのデータマイニングを行う場合においても，意義の少ない見かけ上のルールをフィルタリングする手法として利用できるであろう．さらに，Gene Set Enrichment Analysis（GSEA）[Subramanian 05]等の統計解析にも用いることができる．

ヒト遺伝子の統合データベースである H-InvDB のもとのソースは，GO などのように階層構造や ID，さらに URI が付与されているものもあるが，そうでないものも多く，今回は扱っていないが長い文のようなテキストの記述も多い．これらに ID（URI）をつけ、内容を細かく扱えるようにすすめていくことがセマンティック Web にもとづくデータ統合とその基盤作りに欠かせない．本論文では正の相関のみを扱ったが，負の相関については今後の課題である．本研究の結果は H-InvDB のエンドポイント（http://h-invitational.jp/sparql/）から提供する予定である．

4. おわりに

異なるデータベースで定義されている用語間について，遺伝子をベースとして相関解析を行い，生物学上の関連があるがデータ上は結びつけられていない概念間を同定した。そして、具体な関連する例を検討しつつ、それらを結合するためのプロパティを提案し、付与することによりオントロジーを構築した． 今後これらのリンクを充実させ、データ統合の質が高めていきたい。また概念間の距離（類似度）を付与しているため、これらを積極的に利用した高度な解析や利用がされることを期待している。

参考文献

[山口 11] 山口敦子, 片山俊明: データベース統合利用基盤としてのセマンティックウェブ技術, 細胞工学, 学研メディカル秀潤社, Vol. 30, No. 11, pp.1210-2011. 2011

[Imanishi 04] Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. et al.: Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS biology, Vol. 2, No. 6, pp.e162. 2004

[Takeda 12] Takeda, J., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., Obi, N., Habara, T., Gojobori, T. and Imanishi, T.: H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. Nucleic acids research, Vol. 41, pp.D915-919. 2013

[GO 13] The Gene Ontology Consortium, Gene Ontology annotations and resources. Nucleic Acids Res, 2013. 41(Database issue): p. D530-5.

[Euzenat 13] Euzenat, J. and P. Shvaiko, Ontology Matching2007: Springer-Verlag.

[W3C 09] W3C. *SKOS Simple Knowledge Organization System Reference*. 2009; Available from: http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[North 13] North, J.A., et al., ATP-dependent nucleosome unwrapping catalyzed by human RAD51. Nucleic acids research, 2013. 41(15): p. 7302-12.

[Subramanian 05] Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 2005. 102(43): p. 15545-50.