RG-001

概念階層を用いたタンパク質構造・機能情報に基づく関連文献検索支援システムの構築

A support system for document retrieval based on the hierarchies on protein structure and function information

藤本 亮* 宮西 一徳† 尾崎 知伸‡ 大川 剛直*

Ryo Fujimoto* Kazunori Miyanishi† Tomonobu Ozaki‡ Takenao Ohkawa*

1 はじめに

タンパク質は他の物質と結合，または相互作用することで，機能を発現することが知られている．近年，タンパク質構造解析に関する研究の発展と共に，タンパク質の構造や機能に関わる情報が大量の文献に記載され，電子的に利用可能となっている．一方，アクセス可能な情報が増大しても，情報入手に費やす時間には限りがあるため，それら電子化された文献から，必要な知識だけを集め，整理し提供するシステムが必要となる．しかし，特定のキーワードを用いた全文検索では，利用者が求めている情報を漏らす恐れや，逆に多くの不要な情報までも抽出されるという不都合が起こり得る．更に，タンパク質の構造情報や機能情報が記載されている構造解析文献においては，それぞれがどの程度の関連性を持ち，どのような側面から関連性を持つのかを基に関連文献を探す必要がある．本論文で述べる構造解析文献における関連性とは，文献に記載されているタンパク質に関する情報，すなわちタンパク質が持つ立体構造情報や，発現する機能情報といった観点からの関連性を指す．

これらを背景に，本研究では，現在提供されている様々なバイオ関連データベースを利用して，構造解析文献に含まれている情報と照らし合わせることにより，文献特有の表現の差異を吸収し，また，タンパク質構造や機能が持つ概念階層を利用することで，文献間の構造，機能側面からの関連度を算出する．これにより，検索された文献がどのような側面から，どの程度の関連性を有するのか，という観点からの検索を実現する．

更にそれらを視覚的に捉えることで，関連文献間の比較を容易に行えるように，マップとして表示させるシステムの構築を行った．本システムでは，更に，着目する各タンパク質機能に重みを持たせ，機能側面の必要性に関して利用者自身が操作を行えるようにすることで，実際に必要とされる観点に重きを置いた文献の検索を可能とする．

2 タンパク質情報の取得

2.1 立体構造に関する情報

立体構造とはタンパク質の “形” であり，その “形” は酵素の触媒作用のようなタンパク質機能と密接に関係している．立体構造がどのようになっているかを知ることで，タンパク質間の関連性を測ることが可能となる．タンパク質立体構造データベースとして，PDB[1]が挙げられる [1]．PDB には，分子機能や構造解析実験条件等の情報が文献から抽出されて記述されており，一般に公開されている．つまり，PDB ではタンパク質の構造，機能に関する情報はもちろん，そのタンパク質に関連する文献を参照することも可能である．現在では，解析技術の向上，情報公開の流れに伴い，PDB におけるデータ数が急増しており，主要雑誌でも PDB 登録番号の掲載を義務づけられるようになっている．登録されている各タンパク質は，4 文字の英数字からなる ID コード (PDB_ID) がつけられている．本研究では，PDB を用いて構造解析文献情報を取得する．

2.2 立体構造に基づく分類

実験的手法により決定されたタンパク質立体構造は前述した PDB に登録されている．それら立体構造の類似性や，立体構造から類推される相同性を調べる必要がある場合，SCOP[2]や CATH[3]等の立体構造分類データベースが有用である [2]．これらは，PDB に登録された構造既知のタンパク質ドメインを分類したデータベースである．SCOP では，Class, Fold, SuperFamily, Family, Domain, Spieces の階層ごとにタンパク質の分類が行われている．現在，SCOP が提供するタンパク質立体構造の分類結果は事実上の世界標準とみなされている．タンパク質構造解析を扱う文献においては，構造分類により関連度を比較することが重要であり，それは，ファミリーレベルでは分子機能の観点からも積極的に関連性を主張できる可能性があるからである．タンパク質の類似性において，配列一致度のようなアナログ値ではなく，SCOP の階層という粗いレベルを示した方が趣旨が明確になる場合も多い．どういうドメイン構成であるのか，各ドメインがどのように分類されているか，同じファミリーやスーパーファミリーにはどういったものがあるのかなど，機能に関しても同ファミリー等に分類されたタンパク質からヒントが得られる可能性は大きいと言える．

2.3 機能情報オントロジー

バイオ分野において，遺伝子の名前のつけ方やその解析された機能の記載方法には絶対的な規則はなく，それぞれの研究分野内での慣用的なルールによって決定されている．これは研究者にとって混乱のもとになるばかりでなく，計算機を用いた大規模解析において，重大な錯誤を生む原因となる．Gene Ontology[4]（GO）は，遺伝子の機能を記述するための用語について，生物学分野における共通語彙の作成を目指し，統一化を図るプロジェクトである [3]．主要な目的は，既存の全てのモデル生物種に由来する遺伝子産物について，各々の生物種ごとにデータベース化されているアノテーションの記述間に整合性を与え，知識の対応付けと共有を行うことにある．ゲノム解析の発達により，数万もの遺伝子候補の機能を系統的に記載したり，比較解析を行うための統一的な記載方法が必要不可欠となっている．GO で定義された用語は GO タームと呼ばれ，

*神戸大学大学院工学研究科
†神戸大学大学院自然科学研究科
‡神戸大学自然科学系先端融合研究環

[1] http://www.rcsb.org/pdb/

[2] http://scop.berkeley.edu/

[3] http://www.cathdb.info/

[4] http://www.geneontology.org/

すべての用語が 3 つの概念カテゴリーに分類され，それぞれの機能情報は階層構造で表されている．

GO を利用することで，特有の表現の違いを吸収し，また，階層構造で表された機能情報を用いることで，バイオ関連文献において，キーワード検索のみでは得られない機能関連度を測ることが可能となる．

3 概念階層を用いた関連度算出

2 つの文献間における概念階層を利用した関連性評価について述べる．本研究では，タンパク質構造解析に関する文献を扱うが，これらの文献に記載されている重要な情報として，機能情報と構造情報があげられ，それらは上述したように，それぞれ概念階層を持つ．文献間の関連度を定義するに先立ち，概念階層とその中の 2 概念間の関連度 $d(H,t_1,t_2)$ を以下のように定義する．

$$d(H,t_1,t_2) = \begin{cases} d(H)\times 2 - (P(t_1,c_p(t_1,t_2)) + P(t_2,c_p(t_1,t_2))) & (\exists c_p(t_1,t_2)) \\ 0 & (otherwise) \end{cases} \quad (1)$$

ここで，$d(H)$ は概念階層の深さを表し，$c_p(t_1,t_2)$ は t_1,t_2 における共通の親を，$P(t_1,c_p(t_1,t_2))$ は t_1 と $c_p(t_1,t_2)$ 間の経路の長さを表す．つまり，式 (1) において，数値が大きいほど，関連性が高くなる．図 1 に概念階層を用いた関連性評価の例を示す．例えば，"M/G1 Transition","cell cycle arrest" 間の距離としては，ここでの概念階層の深さは 4 であり，共通の親までのそれぞれの概念の距離の和は 3 となる．そのため，式 (1) において，この概念間距離は"5"となる．後述する GO や SCOP の概念階層における文献間の関連性評価は，この計算式を基に評価を行う．

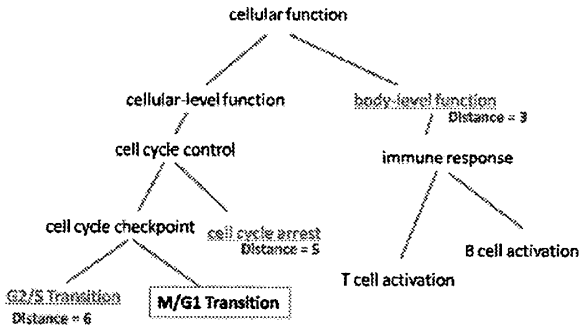

図 1: 概念間関連性の評価例．

3.1 機能階層に基づく関連性の算出

GO における概念間の関係の定義としては「下位の概念に関係付けられている遺伝子産物は，上位の概念においても同様に関係付けられなければならない」という規則に基づいている．まずは PDB により得られる構造情報，機能情報を基に，GO において定義された，それらに対応する機能を取得する．

GO には概念階層における関係性を結びつけるタームとして，「is_a」関係と，「part_of」関係がひとつの概念階層に混在している．「part_of」関係は曖昧性を持つため，概念階層における距離算出には「is_a」関係のみを考慮する．PDB により得られた情報から GO における階層位置を取得し，「is_a」関係を辿ることでタンパク質における機能情報の概念階層を得る．図 2 に「is_a」関係を辿ることによる階層取得の例を示す．

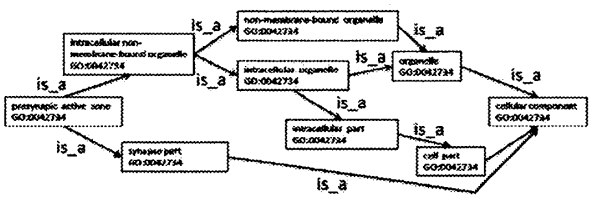

図 2: 「is_a」関係を辿ることによる概念階層の取得．

GO における関連度算出に関しては，入力文献に記載されているタンパク質が持つ各機能に重みを置くことを考えるため，文献に記載されている各機能それぞれと概念階層を対応づけることを考える．また，関連文献に関しては，入力文献における機能それぞれとの関連度を測るために，機能における概念階層は 1 つにまとめる．そして，入力文献の各機能毎に，関連文献が持つ概念階層との比較を行い，機能毎に算出した値の総和を入力文献と関連文献との関連度とする．ここで，式 (1) を用いることにより，GO の概念階層を用いた関連度を以下のように定義する．ここで，$d_{go}(H,D_1,D_2)$ は GO が持つ機能概念階層 H における 2 つの文献 D_1,D_2 間の関連度である．また，ω_{t_1} は文献 D_1 中の t_1 が持つ機能の重みであり，その重みは利用者が与える．

$$d_{GO}(H,D_1,D_2) = \sum_{t_1\in GO(D_1)} (\omega_{t_1} \times \max_{t_2\in GO(D_2)} (d(H,t_1,t_2))) \quad (2)$$

$GO(D_1)$ は，関連文献が持つ機能概念を GO が持つ概念階層に割り当てた概念集合を表す．GO に関しては，この関連性評価を用いることで，入力文献が持つ機能の重みを操作し，重みに応じた関連度を測ることとする．

3.2 構造階層に基づく関連性の算出

SCOP においては先述したように 6 階層の立体構造における分類が存在する．この分類において，各 PDB のタンパク質立体構造情報は基本的にどのドメインに属するかといった位置づけで配置されており，更に細かく種の分類が行われている．

SCOP によるタンパク質の立体構造分類においては，同一タンパク質であっても，タンパク質チェインによって構造分類が多少異なる場合が存在する．つまり，PDB_ID によっては構造分類を複数持つ場合が存在する．しかし，複数の構造分類を有するものであっても，その分類数は少数であるため，構造分類に関しては重みを置くことは考慮しない．つまり，SCOP においては，各構造毎に複数の階層を持たせ，それぞれに関する関連度を算出することは行わない．よって各文献において，それぞれ 1 つの階層しかもたないため，評価式を，以下のように定義する．$d_{scop}(H,D_1,D_2)$ は SCOP が持つ構造概念階層 H における 2 文献 (D_1,D_2) 間の関連度となる．

$$d_{scop}(H,D_1,D_2) = \max_{t_1\in SCOP(D_1),t_2\in SCOP(D_2)} d(H,t_1,t_2) \quad (3)$$

$SCOP(D_1)$ は，関連文献が持つ分類構造すべてを SCOP が持つ概念階層に割り当てた概念集合である．つまり，概念階層における 2 つの概念間に関して最短の 2 概念を選択することで，タンパク質構造分類を行う．得られた結果により，2 文献間において，どのような構造側面からの関連性を持っているかを評価する．

4 システムの構築と評価

4.1 システムの構築と概要

構築した関連文献検索支援システム全体の概要を図 3 に示す．図 3 のように様々なバイオ関連データベースにおけるタンパク質立体構造情報や，それを記載した文献における書誌情報を XML データベース [4, 5] へ取り込み，それを Java サーブレット上で処理し，ブラウザ上で表示させる．

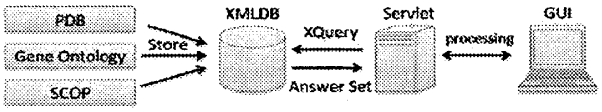

図 3: システム全体の概要．

本システムでは，PDB に登録されている構造解析文献を対象とする．このため，文献に関する情報を持つ PDB_ID が入力となる．PDB_ID の入力を行うと，その ID が持つ書誌情報と，それに対応するタンパク質立体構造分類，機能情報を SCOP と GO から，要約を文献データベース PubMed[6][5]から取得し，ブラウザ上に表示する．また，年代やマップ表示における軸の指定，低関連度の削除等の設定を行うことも出来る．その表示例を図 4 に示す．

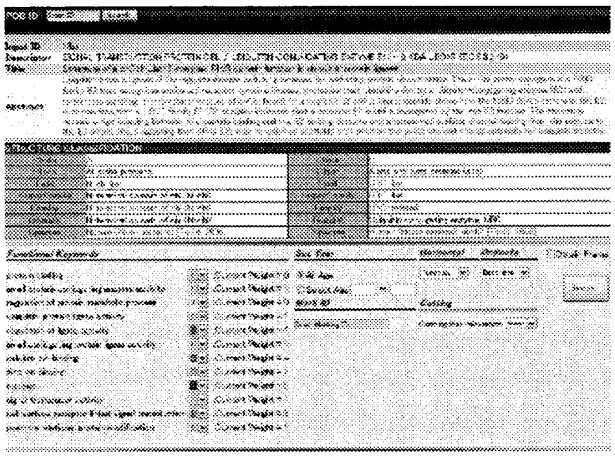

図 4: PDB_ID の入力によって得られる書誌情報，タンパク質情報の表示．

続いて，図 4 に表示されているように，入力された文献中で示されているタンパク質の各機能について重みを指定する．この状態が図 5 に示した状態となる．

図 5 では，3 章で算出した関連度を利用し，横軸を機能関連度，縦軸を出版年，立体構造分類を色によって分類している．また，本システムでは，入力文献以外にも文献を指定することで，その指定された文献をマップ上に強調表示させることができる．そこで，入力文献に対し，関連があることが既にわかっている文献の様な，特に注目すべき文献を指定する．これにより，入力文献に対し関連性があると分かっている既知文献と似通った文献を検索したい場合等に，機能の重みを操作し，その

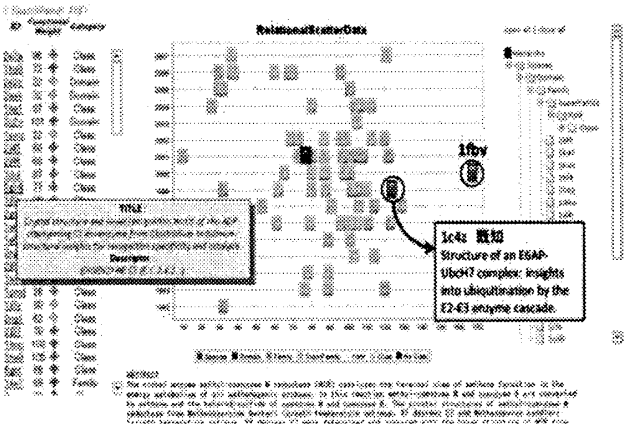

図 5: マップによる関連文献表示．

文献と同様の動きをする文献が存在するならば，それらの文献間の関連性は強いと判断することができる．

4.2 適用実験に基づく評価

タンパク質構造解析文献における関連文献取得における本システムの有用性を示す．適用実験に当たり，PDB_ID「1fbv」，記載されているタンパク質「SIGNAL TRANSDUCTION PROTEIN CBL / UBIQUITIN-CONJUGATING ENZYME E12-18 KDA UBCH7」，文献タイトル「Structure of a c-Cbl-UbcH7 comp lex: RING domain function in ubiquitin-protein ligases」を入力として扱う．この文献に含まれている機能情報としては，図 4 に示されたように，“ubiquitin-protein ligase activity”, “small conjugating protein ligase activity”, “signal transducer activity”, “cell surface receptor linked signal transduction” 等が挙げられる．

ここで，機能情報の重みをすべて均一（ここでは初期状態である重み “1” を指定）に置いた検索結果を図 5 に示す．また，ここでユーザにとって有用であり，入力文献との関連性が強い情報として「1c4z」を与えている．この既知の立体構造に関する文献と関連性が強い文献は，ユーザにとって参照すべき文献である．図 4 で表示された機能キーワード横のチェックボックスを操作し，重みを変えることにより，既知文献と同様の動きを起こす文献を取り上げることで，参照すべき関連文献の取得を行う．このために，この既知文献の関連性が大きくなるように機能の重みを操作することを考える．

図 5 の状態から，上記の機能キーワードにおける “ubiquitin-protein ligase activity” の重みを大きく（ここでは最大値である重み “5” を指定），他の機能キーワードの重みを小さく（“0” もしくは “1”）していくと，関連度が変化し，入力とした文献に対して，いくつかの文献が近づく，もしくは離れるといった状態となる．この「ubiquitin」に関する機能の重みを大きくした状態を図 6 に示す．このように重みによる関連度の変化によって，利用者の意図や目的を反映した関連性を測ることができる．また，ここで既知文献における構造情報として入力した「1c4z」も，この重み付けによって入力文献に近づいてきていることが見て取れる．よって，この重み付けによって「1c4z」と同様の関連度の変化を起こし，関連度が上昇している文献は，「1c4z」と同様参照すべき構造が記載された文献であるといった判断を行うことができる．

[5]http://www.ncbi.nlm.nih.gov/sites/entrez

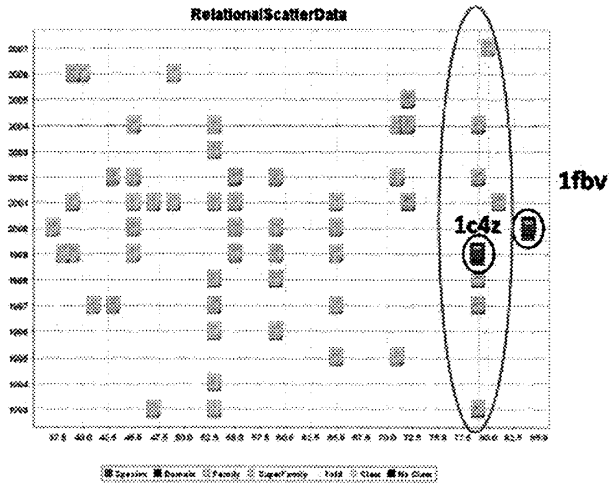

図 6: “ubiquitin” に関する機能の重みを大きく置いたマップ.

表 1: マップにより得られた関連文献の例.

ID	構造分類	タイトル
1c4z	Species	Structure of an E6AP-UbcH7 complex: insights into ubiquitination by th E2-E3 enzyme cascade.
1u9a	Domain	Crystal structure of murine/human Ubc9 provides insight into the variability of the ubiquitin-conjugating system.
2uyz	class	Noncovalent Interaction between Ubc9 and Sumo Promotes Sumo Chain Formation.

表 1 に，上記の操作によって得られた，「ubiquitin」に関して強い関連性を持つ文献の例を示す.

取得された文献は，重み付けされた機能「ubiquitin-protein ligase activity」に関して，タイトル等から，既知文献も含め，関連性が強いものが取得されていると判断できる．またこれらの文献は，「1c4z」と同様の関連度における移動を起こしたため，「1c4z」とも関連性が強い文献であることが判断できる.

また，PDB_ID「2uyz」に記載されている，“Sumo” とは「ユビキチン様タンパク質」と呼ばれるものであり，もちろん「ubiquitin」に対する関連性は存在するものである．このような情報は既存のシステムにおける特定キーワードによる検索では，見つけることは困難である．これに対し，本システムでは，バイオ分野特有の表現の違いを，バイオ関連データベースを利用し，同義語を包含したことで，キーワードとしての一致だけでは取りえない情報に関する関連性も取得することができている.

Google Scholar[6] や PubMed などのような特定のキーワードを用いた既存の（全文）検索システムでは，単に文字列が含まれているだけでも検索結果として表示されてしまう．そのため，検索結果数も多く，更にその出力した文献の羅列となるため，実際に各文献で記述されているタンパク質が，どのような構造を持ち，どのような構造・機能情報を持ち得るのかという側面からの関連性を知ることは容易ではなく，そういった側面から文献を探すことは困難である．対して，本システムでは，本章で示したように，さまざまな機能の重みを操作することで，機能における関連度を視覚的に捉えることができ，文献に記述されている複数の機能観点から関連性を測ることが可能である.

[6]http://scholar.google.co.jp/schhp

このように，本システムでは，膨大な量のタンパク質構造解析文献の中から，タンパク質の立体構造や機能情報を考慮することによる検索を行うことで，構造解析文献特有の側面から，関連文献間の関連度による比較を行い，既存システムでは時間がかかる文献の取得を効率よく行うことが出来る．また本システムでは，これら検索された文献との構造分類的な関わり，他機能との関わりを視覚的に捉えることで，実際に参照すべき文献であるかという判断に関しても，容易に行えることを示した.

5 結論

タンパク質構造解析文献の検索において，公開されているバイオデータベースを用い，立体構造や機能情報の側面から概念階層を利用することで，関連文献との関連性を測り，かつ，それをマップとして表示し視覚的に捉えることで，関連文献の発見を容易にするシステムを構築した．また適用実験を行うことにより，入力文献に対し，その文献に記載されているタンパク質情報を取得し，各々の機能に重みを置くことで，機能関連度からの文献検索の有用性を示した.

本システムにおける今後の課題としては次のことが挙げられる．本システムでは，対象文献として，PDB から参照されている文献のみを扱ったが，それ以外の文献に関しても文献内部を参照することで，タンパク質に関する情報を抽出し，その文献に記載されているバイオデータに関する機能階層や，分類階層を持たせる必要がある．そこで，文献からのタンパク質の情報抽出に関する技術の導入と，その技術と本システムとの統合を今後の課題とする．そのような技術を導入することで，さらに，本システムの有用性が高まると考えられる.

参考文献

[1] Berman, H. M. *et al.*: The Protein Data Bank., *Acta Crystallogr D Biol Crystallogr*, Vol. 58, No. Pt 6 No 1, pp. 899–907 (2002).

[2] Murzin, A. G. *et al.* : SCOP: a structural classification of proteins database for the investigation of sequences and structures., *J. Mol. Biol.*, Vol. 247, pp. 536–540 (1995).

[3] Ashburner, M. *et al.*: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., *Nat Genet*, Vol. 25, No. 1, pp. 25–29 (2000).

[4] 山田祥寛：XML データベース入門 NeoCore/Xpriori で XMLDB を極める, 翔泳社 (2006).

[5] Chamberlin, D.:XQuery: An XML query language., *IBM Syst. J.*, Vol. 41, No. 4, pp. 597–615 (2002)

[6] McEntyre, J. Lipma, D.:PubMed: bridging the information gap., *Cmaj.*, Vol. 164, pp. 1317–1319 (2001).