■S2 群（ナノ・量子・バイオ）- 6 編（バイオインフォマティックス）

4 章 データベース・ツール

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

■概要■

塩基配列，遺伝子発現，DNA メチル化などの網羅的な測定技術の発展に伴い，大規模なデータベース化が進んでいる．本章では現在構築されている様々なデータベースについて概説する．

【本章の構成】

本章では，バイオインフォマティクスで利用されるデータベースとツールについて説明する．4-1 節では，様々なデータベースを統合した統合データベースについて説明する．4-2 節では，塩基配列に関するデータベースについて説明する．4-3 節では，塩基配列の一部が変化した多型に関するデータベースについて説明する．4-4 節では，タンパク質に関するデータベースについて説明する．4-5 節では，様々な研究者が独立した発見したが実際は同じであるような遺伝子を統一的に取り扱うための概念であるオントロジーについて説明する．4-6 節では遺伝子同士の関係（ある遺伝子が別の遺伝子の発現を促進あるいは抑制する）について記述したパスウェイに関するデータベースについて説明する．4-7 節では，発現データに関するデータベースについて説明する．4-8 節では，文献情報に関するデータベースについて説明する．4-9 節ではゲノムワイド関連研究について説明する．4-10 節では，バイオインフォマティクスで用いられる各種プログラミング言語について説明する．

■S2 群 - 6 編 - 4 章

4-1 統合データベース

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

ライフサイエンスにおいて，データベースの作成は研究の副産物ではなく，研究と同等に価値のあるものである．データをどのような方法で収集するか，得られたデータに対してどのような意味付けを行うか，意味付けされたデータをほかのデータとどのように関連付けるかが重要な項目となってくる．

また，単にデータを収集するだけでなく，如何にして関連性のあるデータを効率的に検索できるか，発見された関連性が新たな知見を与えるかなどデータを取り扱う技術を発展させるためにもデータベースの構築には様々な知識・技術が必要とされる．

現在までに様々な分野のデータベースが個別に発展してきたが，今後のライフサイエンスの発展を考えると，今までのデータベースから得られた知識・知見を統合して活用していくことが必要不可欠となる．

データベースの統合には，データを共有するためのルールや，半構造的なデータを管理するシステムの開発などが必要となる．また，新たな知識の発見を支援するためのオントロジー技術や統一的なインタフェースの開発など，データの構造・データの利用の両面からデータベースを統合していくことが必要となる．

このような考えのもと，現在までにいくつかの統合データベースが作成されている．代表的な生命科学系の統合データベースを表 4.1 に示す．

表 4.1　代表的な生命化学系統合データベース

名称	略号	所轄
バイオサイエンスデータベースセンター	NBDC[1)]	文部科学省
国立研究開発法人 医薬基盤・健康・栄養研究所	NIBIO[2)]	厚生労働省
農畜産物ゲノム情報データベース	AgrID[3)]	農林水産省
ライフサイエンス統合データベース	MEDALS[4)]	経済産業省

また，これらの生命科学系データベースの更なる統合を目指して，データベースのカタログ，横断検索，アーカイブ構築などの連携が上記の 4 つのデータベース間で進められている[5)]．

（1）NBDC の概要

NDBC では，上記の問題を解決するために，ライフサイエンス分野のデータベースを統合し，世界のユーザに貢献できる日本のデータベースセンターとなることを目的としており，NBDC の持つヒトデータベースや，後述する塩基データベース，タンパク質データベース，パスウエイデータベースなどへのポータルサイトとしての役割も果たしている．

（2）NIBIO の概要

創薬支援データベース，難病研究資源バンクデータベース，薬用植物データベース，メディカル・バイオリソースデータベースなどの各種データベースへのポータルサイトとしての役割を果たしている．

（3）AgrID の概要

イネなどの作物，動物，カイコなどの昆虫の農畜産物ゲノム情報データベースの統合データベースとしての機能のほかに，次世代シーケンサーを用いた農畜産物の高次解析システム（Galaxy），及び，Galaxy から得られた大量の塩基配列データを保存・解析するためのデータベースシステム SOGO を提供している．

（4）MEDALS の概要

様々な研究機関が提供する DNA・ゲノム，タンパク質，RNA などのデータベースや解析ツールへのポータルサイトとしての役割だけでなく，それぞれのデータベースやツールの解説も作成されており，データベースやツールの概略を理解するためのデータベースとしても機能している．

■参考文献

1) https://biosciencedbc.jp/
2) http://www.nibiohn.go.jp/nibio/data/
3) http://agrid.dna.affrc.go.jp/
4) https://medals.jp/
5) http://integbio.jp/ja/

■S2 群 - 6 編 - 4 章

4-2 塩基配列，遺伝子

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

ヒトゲノムの解析が終了し，全塩基配列が決定された．塩基配列，遺伝子に関するデータベースも数多く作成されている．表 4.1 にヒト遺伝子の塩基配列情報を取り扱う代表的なデータベースを示す．

表 4.1　代表的な塩基配列データベース

名称	略称	所属
DNA Data Bank of Japan	DDBJ[1, 2]	日本
The European Molecular Biology Laboratory - The European Bioinformatics Institute	EMBL-EBI[3, 4]	欧州
National Center for Biotechnology Information	NCBI[5]	米国

これらのデータベースは，国際塩基配列データベース International Nucleotide Sequence Database Collaboration（INSDC）[6, 7]により，情報交換が行われてデータベースの整合性を保っている．登録されている塩基配列データの構成[8]を表 4.2 に示す．

表 4.2　登録データ種別

<table>
<tr><td rowspan="3">実配列データ</td><td rowspan="2">シークエンサ出力データ</td><td>次世代シークエンサデータベース</td></tr>
<tr><td>シングルパスリードデータベース</td></tr>
<tr><td>Annotated/Assembled データ</td><td>塩基配列データ</td></tr>
<tr><td rowspan="2">研究プロジェクト・サンプル</td><td>BioProject</td><td>研究プロジェクトとプロジェクトに由来するデータをまとめたデータベース</td></tr>
<tr><td>BioSample</td><td>実験データを得るために使用された試料についての情報を管理するデータベース</td></tr>
<tr><td>制限アクセス公開</td><td colspan="2">個人に由来する遺伝学的データと匿名化された表現型情報のデータベース</td></tr>
</table>

また，米国の NCBI では遺伝子を EntrezID で，欧州の EMBL-EBI では遺伝子及び転写情報を EnsemblID で管理している．

また，代表的な RNA データベースとして RNA interference（RNAi）のデータを取り扱う GenomeRNAi[9, 10]，micro RNA（miRNA）のデータを取り扱う miRBase[11, 12]がある．

■参考文献

1)　http://www.ddbj.nig.ac.jp/index-j.html

2) J. Mashima, Y. Kodama, T. Fujisawa, T. Katayama, Y. Okuda, E. Kaminuma, O. Ogasawara, K. Okubo, Y. Nakamura, and T. Takagi：“ DNA Data Bank of Japan, ” Nucleic Acids Res., 45(D1), pp.D25-D31, 2017.
3) http://www.ebi.ac.uk/
4) D.R. Zerbino, P. Achuthan, W. Akanni, M.R. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, C.G. Girón, L. Gil, L. Gordon, L. Haggerty, E. Haskell, T. Hourlier, O.G. Izuogu, S.H. Janacek, T Juettemann, J.K. To, M.R. Laird, I. Lavidas, Z. Liu, J.E. Loveland, T. Maurel, W. McLaren, B. Moore, J. Mudge, D.N. Murphy, V. Newman, M. Nuhn, D. Ogeh, C.K. Ong, A. Parker, M. Patricio, H.S. Riat, H. Schuilenburg, D. Sheppard, H. Sparrow , K. Taylor, A. Thormann, A. Vullo, B. Walts, A. Zadissa, A. Frankish, S.E. Hunt, M. Kostadima, N. Langridge, F.J. Martin, M. Muffato, E. Perry, M. Ruffier, D.M. Staines, S.J. Trevanion, B.L. Aken, F. Cunningham, A. Yates, and P. Flicek：“ Ensembl 2018, ” Nucleic Acids Res., 46(D1), pp.D754-D761, 2017.
5) https://www.ncbi.nlm.nih.gov/
6) http://www.insdc.org
7) M. Blaxter, A. Danchin, B. Savakis, K.F. Kobayashi, K. Kurokawa, S. Sugano, R.J. Roberts, S.L. Salzberg, and C. Wu：“ Reminder to deposit DNA sequences, ” Science, 352(6287), pp.780, 2016.
8) http://www.ddbj.nig.ac.jp/sub/data_categories-j.html
9) http://www.genomernai.org/
10) E.E. Schmidt, O. Pelz, S. Buhlmann, G. Kerr, T. Horn, and M. Boutros：“ GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update, ” Nucleic Acids Res., 41(D1), pp.D1021-D1026, 2013.
11) http://www.mirbase.org/
12) A. Kozomara and S. Griffiths-Jones：“ miRBase: annotating high confidence microRNAs using deep sequencing data, ” Nucleic Acids Res., 42(D1), pp.D68-D73, 2014.

■S2 群 - 6 編 - 4 章

4-3 多　　型

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

生物学と医学において，遺伝子型と表現型との関係を理解することは中心的な目標の一つである．遺伝子型とは遺伝子の塩基配列や，染色体上での対立遺伝子の優性，劣性の関係といった遺伝子の基本構成をいう．表現型とは遺伝子型の持つ情報に基づいて，実際に発現され外部から観測できる形質をいう．遺伝子多型の系統的な研究には，すべての対立遺伝子において，変位の配列と，その変位が発生する頻度情報が必要となる．国際 HapMap プロジェクト（1000 ゲノムプロジェクト）[1]は，アフリカ（AFR），東アジア（EAS），ヨーロッパ（EUR），南アジア（SAS），アメリカ（AMR）の 5 つの主要な地域からなる 14 のグループに対して，350 万個の SNP の対立遺伝子の頻度と，変異体の間の相関パターン（連鎖不均衡，LD）を調べ，データベース化した．

1000 ゲノムプロジェクトの最終フェーズ（Phase 3）[2]では，グループの数を 26 に増やし，全ゲノム配列決定と標的化エクソーム配列決定の 2 種類の方法を用いてすべての個体の配列を決定している．更に，高密度 SNP マイクロアレイを用いて個体及び利用可能な一次親族（主に成人）の遺伝子型を決定した．この結果から個々の遺伝子型及びハプロタイプを推定することが可能となっている．ハプロタイプとは，同じ染色体上にある，すなわち，同じ親から継承される対立遺伝子の組合せのことをいう．表 4.1 に 1000 ゲノムプロジェクトの Phase 1 及び Phase 3 の固体グループ名と固体数の一覧を示す．

■参考文献

1) The 1000 Genomes Project Consortium：“ A map of human genome variation from population-scale sequencing, ” Nature, 467, pp.1061-1073, 2010.
2) The 1000 Genomes Project Consortium：“ A global reference for human genetic variation, ”Nature, 526, pp.68-74, 2015.

表 4.1　1000 ゲノムプロジェクトのグループ

集団	コード名	パネル名	固体数 (Phase1)	固体数 (Phase3)
アフリカ				
ナイジェリア・エサン族	ESN	AFR		99
ガンビア・マンディンカ族	GWD	AFR		113
ケニア・ルイヤ族	LWK	AFR	97	99
シエラレオネ・メンデ族	MSL	AFR		85
ナイジェリア・ヨルバ族	YRI	AFR	88	108
バルバドス・アフリカ系カリビア	ACB	AFR/AMR		96
南西アメリカ・アフリカ系アメリカ	ASW	AFR/AMR	61	61
アメリカ				
コロンビア	CLM	AMR	60	94
ロサンゼルス・メキシコ系アメリカ	MXL	AMR	60	64
ペルー	PEL	AMR		85
プエルトリコ	PUR	AMR	55	104
東アジア				
中国・シーサンパンナ・タイ族	CDX	EAS		93
北京・漢民族	CHB	EAS	97	103
南漢民族	CHS	EAS	100	105
東京・日本民族	JPT	EAS	89	104
ベトナム・キン族	KHV	EAS		99
ヨーロッパ				
北西ヨーロッパの祖先・ユタ州の住民	CEU	EUR	85	99
イングランド・スコットランドのイギリス人	BGR	EUR	89	91
フィンランド	FIN	EUR	93	99
スペイン・イベリア	IBS	EUR	14	107
イタリア・トスカニ	TSI	EUR	98	107
南アジア				
バングラデシュ・ベンガル	BEB	SAS		86
テキサス・グジャラート州・インド	GIH	SAS		103
イギリス・テルグ・インド	ITU	SAS		102
パキスタン・パンジャブ	PJL	SAS		96
イギリス・スリランカ・タミル	STU	SAS		102
合　計			1092	2504

■S2 群 - 6 編 - 4 章

4-4 タンパク質

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

表 4.1 に国際的に統一化されたタンパク質構造情報を取り扱う代表的なデータベースを示す.

表 4.1　代表的なタンパク質データベース

名称	略称	所属
Protein Data Bank Japan	PDBj[1, 2)]	日本
Protein Data Bank in Europe	PDBe[3, 4)]	欧州
Protein Data Bank	PDB[5, 6)]	米国

PDB で取り扱うタンパク質データベース情報としては，対象であるタンパク質の分子情報や，各原子の三次元座標，構造の特色（分子全体，二次構造，…），化合物データ，文献情報などがある．化合物検索や，アミノ酸の配列検索，タンパク質の構造検索などの様々な検索サービスが存在する．また，タンパク質，ペプチド，塩基配列などの核磁気共鳴スペクトルスコピーのデータベースとして BMRB[7, 8)]がある．PDBj，PDBe，PDB は，生体高分子の立体構造を国際的に統一したデータベースとして運営されている．

■参考文献

1) https://pdbj.org/
2) A.R. Kinjo, G.J. Bekker, H. Suzuki, Y. Tsuchiya, T. Kawabata, Y. Ikegawa, and H. Nakamura：“ Protein Data Bank Japan (PDBj): Updated user interfaces, Resource Description Framework, analysis tools for large structures, ” Nucleic Acids Res., 45(D1), pp.D282-D288, 2017.
3) http://www.ebi.ac.uk/pdbe/
4) S. Mir, Y. Alhroub, S. Anyango, D.R. Armstrong, J.M. Berrisford, A.R. Clark, M.J. Conroy, J.M. Dana, M. Deshpande, D. Gupta, A. Gutmanas, P. Haslam, L. Mak, A. Mukhopadhyay, N. Nadzirin, T. Paysan-Lafosse, D. Sehnal, S. Sen, O.S. Smart, M. Varadi, G.J. Kleywegt, and S. Velankar：“ PDBe: towards reusable data delivery infrastructure at protein data bank in Europe, ”Nucleic Acids Res., 46(D1), pp.D486-D492, 2018.
5) http://www.rcsb.org/pdb/
6) S.K. Burley, H.M. Berman, C. Christie, J.M. Duarte, Z. Feng, J. Westbrook, J. Young, and C. Zardecki：“ RCSB Protein Data Bank: Sustaining a living digital data resource that enables breakthroughs in scientific research and biomedical education, ” Protein Sci., 27(1), pp.316-330, 2018.
7) http://www.bmrb.wisc.edu/
8) E.L. Ulrich, H. Akutsu, J.F. Doreleijers, Y. Harano, Y.E. Ioannidis, J. Lin; M. Livny, S. Mading, D. Maziuk, Z. Miller, E. Nakatani, C.F. Schulte, D.E. Tolmie, R.K. Wenger, H. Yao, and J.L. Markley：“ BioMagResBank, ” Nucleic Acids Res., 36(D1), pp.D402-D408, 2008.

■S2 群 - 6 編 - 4 章

4-5 オントロジー

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

4-5-1 オントロジー

オントロジーとは元来は哲学の用語であり，人間の知識を扱う分野・手法のことである．コンピュータを用いた知識表現としてのオントロジーにおいては，知識に関する体系的な理論や，概念化の明示的な規約・使用といった意味で使用されることが多い．つまり，ある分野の知識を計算機で処理可能とするために，明示的かつ論理的に記述し，知識の共有・再利用を可能とするものとして用いられる．

対象として生物を理解し，モデルとして体系化した場合を考えてみる（生物のオントロジー）．生物は大きく原核生物と真核生物に分類される．真核生物は更に動物，植物，菌類，原生生物といった生物に分類される．動物は更に脊索動物，半索動物など細かく分類される．このように生物のオントロジーでは徐々に複雑になっていく進化の系統樹のような分類となることが分かる．

4-5-2 ジーン・オントロジー

上述したように対象を理解し，モデルとして体系化することがオントロジーの構築である．この対象を遺伝子とその機能としたものがジーン・オントロジー（Gene Ontology）である．

近年の DNA シーケンサや DNA マイクロアレイ技術の急速な発展により，膨大な量の遺伝子関連情報（遺伝子名やその機能など）がデータベース化されている．しかし，同一の遺伝子が複数の機能を併せ持つことは多く，これらの機能が別々の研究者により異なった時期に研究され発表されることも多い．そのため，実際は同一の遺伝子であるにも関わらず，当初は別々の遺伝子・タンパク質と考えられていたため，複数の別名（Synonym, Alias）を持つものもが数多く存在する．例えば，がん抑制遺伝子として有名な P53 は，複数の Synonym を持っている（BCC7，LFS1，TRP53）．また，検索の際に主たるキーとなる遺伝子名がどのデータベースでも同じであるとは限らないことも多い．このように複雑な形で存在する遺伝子関連情報を有効に活用するためには，遺伝子情報をオントロジー化して記述していく必要がある．オントロジー化された情報を用いることにより，異なるデータベース間での，データの統合や比較を行うことが可能になる．

遺伝子に関するオントロジー情報として公開されているものに Gene Ontology Database（GO）[1, 2]がある．GO では遺伝子を以下の 3 つのカテゴリーに分類する．

- 生物学的プロセス（Biological Process）
- 細胞の構成要素（Cellular Component）
- 分子機能（Molecular Function）

このような GO 情報を用いてデータ解析を対話的に行う WEB ベースの解析ツールとして The Database for Annotation, Visualization and Integrated Discovery（DAVID）[3, 4]がある．

DAVID ではマイクロアレイ発現データを主な解析対象としており，遺伝子を機能別のグループに分類するなどの解析を行うことができる．

似たような機能を持つ遺伝子は似たような発現傾向を示すことがある．そこでマイクロアレイ発現データなどにより似たような発現パターンを示す遺伝子グループを抽出したときに，これらの遺伝子が機能的にどのようなグループに属するのかを解析したいときなどに，この DAVID が使われることが多い．

また，DAVID では，遺伝子 ID として，遺伝子名やその Synonym だけでなく，上述した NCBI Entrez ID や EMBC Ensembl ID，更にはマイクロアレイの Probe ID などでも解析が行えるように作成されている．また，この機能を用いて，遺伝子 ID の相互変換を行うことも可能となっている．

■参考文献

1) http://www.geneontology.org/
2) The Gene Ontology Consortium：“ Expansion of the Gene Ontology knowledgebase and resources, ” Nucleic Acids Res, 45(D1), pp.D331-D338, 2017.
3) http://david.abcc.ncifcrf.gov/
4) D.W. Huang, B.T. Sherman, and R.A. Lempicki：“ Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources, ” Nature Protoc., 4(1), pp.44-57, 2009.

■S2 群 - 6 編 - 4 章

4-6 パスウエイ

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

4-6-1 KEGG

Kyoto Encyclopedia of Genes and Genomes（KEGG）は，細胞，組織，生態系といった生物システムの高次機能をゲノムや分子情報から理解するためのデータベースである[1, 2]．

このデータベースは，遺伝子とタンパク質（ゲノム情報），化学物質（化学情報），相互作用，反応及び関係のネットワーク（システム情報），更に疾患及び薬物情報（健康情報）からなる．

(1) ゲノム情報の概要

ゲノム情報は，生物情報（KEGG GENOME），遺伝子情報（KEGG GENE），配列情報（KEGG SSDB）からなる．

KEGG GENOME

KEGG GENOME は 3 文字あるいは 4 文字の生物コードで特定される生物情報のデータベースであり，外部の配列情報データベースへの参照情報を持つ．データベースにある生物グループを表 4.1 に示す．

表 4.1 KEGG 生物

ドメイン	界	門	綱
			哺乳類
			鳥類
			爬虫類
			両生類
			魚類
		節足動物	昆虫
		線虫	
		軟体動物	
	動物	刺胞動物	
		真正双子葉類	
		単子葉植物	
		緑藻類	
	植物	紅藻類	
	菌類		
真核生物	原生生物		
	細菌		
原核生物	古細菌		

例えば，ヒトは哺乳類に含まれ hsa の 3 文字の生物コードで特定される．

KEGG GENES

公的に利用可能なデータベース（主に NCBI RefSeq[3, 4] と GenBank[5, 6]）から生成された遺伝子データベースである．

KEGG SSDB（Sequence Similarity DataBase）

タンパク質に翻訳される遺伝子間のアミノ酸の配列の相似性情報に関するデータベースである．保存された遺伝子クラスタを検索するのに有効なデータベースである．

(2) 化学情報の概要

化学情報は，生命に関係する化学物質とその反応に関するデータベースであり，化合物構造（KEGG COMPOUND），糖鎖構造（KEGG GLYCAN），化学反応（KEGG REACTION）から構成される．

KEGG COMPOUND

生物学的システムに関連する小分子，生物高分子，化学物質のデータベースである．

KEGG GLYCAN

実験的に決定された糖鎖構造のデータベースであり，CarbBank[7]から得られた構造，論文から得られた構造，KEGG PATHWAY（後述）から得られた構造を含む．

KEGG REACTION

KEGG PATHWAY（後述）に含まれる主に酵素反応に関する化学反応のデータベースである．

(3) システム情報の概要

システム情報は，パスウエイ情報（KEGG PATHWAY），機能階層情報（KEGG BRITE），モジュール情報（KEGG MODULE）から構成される．

KEGG PATHWAY

分子間相互作用や遺伝子ネットワーク間の関係や反応についての文献知識に基づき，機械的でなく人間により作成されたグラフィカルなネットワーク図である．

パスウエイは，代謝（炭水化物代謝，脂質代謝，エネルギー代謝），遺伝子情報（転写，翻訳，タンパク質立体構造の折りたたみ），環境情報（膜輸送，シグナル伝達），分子プロセス（細胞の成長，細胞間の通信），生物システム（免疫システム，細胞分裂システム），疾病（がん，免疫疾病，循環器疾病），医薬品開発の 7 つのグループからなる．

KEGG BRITE

KEGG データベースで取り扱う様々な生物学的物体（遺伝子，化学物質，酵素，糖鎖）の機能的階層構造のデータベースである．KEGG PATHWAY と同様に文献知識に基づき，機械的でなく人間により作成された階層構造テキストである．

(4) 健康情報の概要

健康情報はヒト疾病情報（KEGG DISEASE），薬物情報（KEGG DRUG），環境情報（KEGG ENVIRON），医療情報（KEGG MEDICUS）から構成される．

KEGG DISEASE

遺伝的及び環境的摂動に関する知識を取り込んだヒト疾病に関するデータベースである．KEGG データベースでは，疾病は分子システムの摂動とみなしている．

KEGG DRUG

KEGG データベースでは，薬物は分子システムの摂動剤とみなしている．日本，欧州及び米国で認可された薬物情報の網羅的なデータベースである．情報は薬物の化学構造，化学成分，標的，代謝酵素，及び，ほかの分子相互作用ネットワーク情報を含んでいる．また，日本の市販薬は，処方薬だけでなく一般用医薬品も含んでおり，添付文書の情報も取り込まれている．

KEGG ENVIRON

主に植物の天然産物からなる生薬，精油，及びその他の健康促進物質に関するデータベースである．

KEGG MEDICUS

KEGG データベースは，細胞レベル及び生物レベルの機能を理解するための知識データベースであるとともに，研究成果を実用的なものとし，社会が疾病や薬物の科学的根拠をより理解できるような研究を支援するというゲノム社会革命を目指している．

KEGG MEDICUS は，このような理念のもとに作成された疾病，薬物，健康関連物質の総合データベースであり，前述したライフサイエンス統合データベース（NBDC）の支援を受けている．

4-6-2 LINCS プロジェクト

LINCS プロジェクトは、米国の National Institute of Health（NIH）が提供する，細胞の挙動をネットワークベースで統合したライブラリである（Library of Integrated Network-Based Cellular Signatures：LINCS）[8, 9]．このライブラリは，細胞が様々な遺伝的及び環境ストレスにどのように反応するかを示すデータを公開することによって，パスウェイの詳細な理解や，疾患によって変動したパスウェイを正常状態に戻す治療法を開発することを支援するために作成された．

■参考文献

1) http://www.genome.jp/kegg/
2) M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima：“ KEGG: new perspectives on genomes, pathways, diseases and drugs, ” Nucleic Acids Res., 45(D1), pp.D353-D361, 2017.
3) https://www.ncbi.nlm.nih.gov/refseq/
4) N.A. O’Leary, M.W. Wright, J.R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C.M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V.S. Joardar, V.K. Kodali, W. Li, D. Maglott, P. Masterson, K.M. McGarvey, M.R. Murphy, K. O’Neill, S. Pujar, S.H. Rangwala, D. Rausch, L.D. Riddick, C. Schoch, A. Shkeda, S.S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R.E. Tully, A.R. Vatsan, C. Wallin, D. Webb, W. Wu, M.J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T.D. Murphy, and K.D. Pruitt：“ Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, ” Nucleic Acids Res., 44(D1), pp.D733-D745, 2016.
5) https://www.ncbi.nlm.nih.gov/genbank/
6) D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and E.W. Sayers：“ GenBank, ” Nucleic Acids Res., 41, pp.D36-D42, 2012.
7) S. Doubet and P. Albersheim, “ CarbBank, ” Glycobiology, 2(6), pp.505, 1992.
8) http://www.lincsproject.org/
9) A.B. Keenan, S.L. Jenkins, K.M. Jagodnik, S. Koplev, E. He, D. Torre Z. Wang, A.B. Dohlman, M.C. Silverstein, A. Lachmann, M.V. Kuleshov, A. Ma’ayan, V. Stathias, R. Terryn, D. Cooper, M. Forlin, A. Koleti, D. Vidovic, C. Chung, S.C. Schürer, J. Vasiliauskas, M. Pilarczyk, B. Shamsaei, M. Fazel, Y. Ren, W. Niu, N.A. Clark, S. White, M. Mahi, L. Zhang, M. Kouril, J.F. Reichard, S. Sivaganesan, M. Medvedovic, J. Meller, R.J. Koch, M.R. Birtwistle, R. Iyengar, E.A. Sobie, E.U. Azeloglu, J. Kaye, J.

Osterloh, K. Haston, J. Kalra, S. Finkbiener, J. Li, P. Milani, M. Adam, R. Escalante-Chong, K. Sachs, A. Lenail, D. Ramamoorthy, E. Fraenkel, G. Daigle, U. Hussain, A. Coye, J. Rothstein, D. Sareen, L. Ornelas, M. Banuelos, B. Mandefro, R. Ho, C.N. Svendsen, R.G. Lim, J. Stocksdale, M.S. Casale, T.G. Thompson, J. Wu, L.M. Thompson, V. Dardov, V. Venkatraman, A. Matlock, J. E. Van Eyk, J. D. Jaffe, M. Papanastasiou, A. Subramanian, T.R. Golub, S.D. Erickson, M. Fallahi-Sichani, M. Hafner, N.S. Gray, J.R. Lin, C.E. Mills, J.L. Muhlich, M. Niepel, C.E. Shamu, E.H. Williams, D. Wrobel, P.K. Sorger, L.M. Heiser, J.W. Gray, J.E. Korkola, G.B. Mills, M. LaBarge, H.S. Feiler, M.A. Dane, E. Bucher, M. Nederlof, D. Sudar, S. Gross, D.F. Kilburn, R. Smith, K. Devlin, R. Margolis, L. Derr, A. Lee, and A. Pillai：“ The Library of Integrated Network-Based Cellular Signatures NIH program: System-level cataloging of human cells response to perturbations, ” Cell Systems, pii: S2405-4712(17): pp.30490-30498, 2017.

■S2 群 - 6 編 - 4 章

4-7 発現データ

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

4-7-1 GEO

Gene Expression Omnibus（GEO）[1, 2]は，マイクロアレイ，次世代シーケンシング，研究コミュニティによって提出されたゲノミクスデータをまとめた公的に利用可能なデータベースである．GEO の目的は以下の 3 つである．

1. ハイスループットな機能ゲノムデータを効率的に収集し，堅牢で汎用性のあるデータベースを提供する．
2. 研究コミュニティからの完全で注釈がついたデータを受け入れるための，簡単な手続きを提供する．
3. 対象となる研究及び遺伝子発現プロファイルへの照会，検索，ダウンロードを可能とするユーザフレンドリーなインタフェースを提供する．

GEO で受託するデータは以下の 4 種類である．

1. マイクロアレイ発現データ（Affymetrix 社，Aglient 社，Nimblegen 社，Illumina 社，その他）
2. RT-PCR 発現データ
3. ハイスループットシーケンスデータ
4. 従来型の SAGE データ

GEO ではデータ提供者からプラットフォーム情報，試料情報，系列情報を受託し，これを整理して GEO データセット，GEO プロファイルの形式でデータベース利用者に提供する．

(1) **プラットフォーム情報**

プラットフォーム情報は，マイクロアレイあるいはシーケンサの概要説明と，マイクロアレイのデータ表から構成される．各プラットフォーム情報は固有の GEO 登録番号（GPL***）が割り当てられて管理されている．プラットフォーム情報を用いることで，同一のマイクロアレイに対して，複数の研究者から提出された試料情報をまとめて参照することが可能となる．

(2) **試料情報**

試料情報は，各試料が処理された条件・操作に関する記述，及び試料から得られた測定値が含まれている．各試料情報は固有の GEO 受託番号（GSM***などの番号）が割り当てられて管理されている．試料が参照することができるプラットフォーム情報は 1 つのみであるが，複数の系列情報から参照されることがある．

(3) **系列情報**

系列情報は，関連する試料をグループ化し，研究全体の説明を記述する．試料情報には抽

出されたデータ，結論の要約，試料の分析に関する記述が含まれることもある．各系列情報は固有の GEO 受託番号（GSE***などの番号）が割り当てられて管理されている．

(4) **GEO データセット**

既に説明したように，系列情報は，実験を要約した提出者から提供された情報である．これらのデータを基に，GEO により整理された GEO データセットを構築する．構築された GEO データセットは，生物学的及び統計的に比較可能な GEO 試料情報を整理したものであり，データだけでなく GEO の一連のデータ表示及び分析ツールが利用可能となっている．GEO データセットに含まれる試料はすべて同じプラットフォームのものであるので，データセット内の試料の測定値のバックグラウンド処理や正規化などの事項はデータセット全体で一貫している．

(5) **GEO プロファイル**

GEO プロファイルのデータは整理された GEO データセットから得られた遺伝子発現プロファイルである．そのため，特定の遺伝子の発現レベルの変化を GEO データセット内にあるすべての試料にわたって調べることが可能となる．GEO プロファイルでは実験条件に関する情報も整理されているので，異なる実験条件で遺伝子の発現がどのように変化するかを容易に確認することができる．また，同様の挙動をする遺伝子に関する GEO 内データベースへのリンクや，関連する情報への外部データベースへのリンクなど，多数のリンク情報も含んでいる．

4-7-2 TCGA

The Cancer Genome Atlas（TCGA）[3, 4]は，National Cancer Institute （NCI）と National Human Genome Research Institute（NHGRI）との共同研究であり，33 種類のがんにおける重要なゲノム変化の包括的かつ多次元マップを持つ．

2 ペタバイト以上のゲノムデータを含む TCGA データセットが公開されており，このゲノム情報はがんの予防，診断及び治療を改善するために利用されている．

実際のデータはポータルサイトである GDC Data Portal（https://portal.gdc.cancer.gov/）から入手可能である．

4-7-3 ArrayExpress

ArrayExpress[5, 6]は，The European Bioinformatics Institute（EMBL-EBI）が提供する，マイクロアレイ及びシーケンシングプラットフォームからの機能的ゲノミクスデータを保存・管理し，再現性のある研究をサポートするためのリポジトリーの一つである．

ArrayExpress では，そのため，マイクロアレイ実験に関する最小情報（MIAME），及びシーケンシング実験に関する最小情報（MINSEQE）のガイドラインに従って情報を提供する．

取り扱うデータは, 直接 ArrayExpress に提出されたものと, NCBI Gene Expression Omnibus（GEO）データベースから提供されたものになる．

4-7-4 ENCODE

ENCODE[7, 8]（Encyclopedia of DNA Elements）コンソーシアムは, National Human Genome Research Institute（NHGRI）が資金を提供する国際共同研究グループである．ENCODE の

目的は，タンパク質及び RNA レベルで作用する要素，細胞及び環境を制御する調節要素を含む，ヒトゲノムにおける機能要素の包括的なリストを構築することである．

ENCODE では，セルライン，組織，全臓器，*in vitro* セルラインなどの様々な生体試料に関して，ChIP シーケンス（ChIP-Seq）により転写因子及びほかのタンパク質のためのゲノム全体の DNA 結合部位を同定や，DNA 結合，RNA 結合，転写，DNA メチル化，ヒストン修飾などの様々なアッセイ結果を網羅したデータベースである．

4-7-5 接続マップ

ヒトゲノムの配列決定により，疾患の遺伝的な基盤への新しい知見が得られた．疾患に関連する遺伝子の同定は行われているが，多くの場合，それらの遺伝子の機能は不明瞭なままであるという問題がある．

同時に，化学生物学と創薬へのアプローチも劇的に拡大してきた．新しい種類の化学ライブラリが作成され，強力なスクリーニング方法が開発され，新しい治療法が臨床の現場で試されている．しかし，特定の化合物の細胞への効果を体系的に決定する方法はまだなく，臨床応用を制限するような化合物の予期せぬ副作用は，薬剤開発プロセスの最終的な段階で発見されることが多い．このような問題に対する可能性のある解決方法としては，遺伝的な摂動（タンパク質機能の反映）及び薬理的摂動（小分子機能の反映）を伴う系統的な摂動を表す包括的な細胞活動のカタログの作成が必要であると考えられる．

類似性が高い細胞活動は有用ではあるが，これまでは確認できていなかった新しい細胞 – 小分子間の接続を表す可能性がある．例えば同じパスウェイ，小分子とそのタンパク質の標的との間の接続や，同様の機能を持つ 2 つの小分子の間で構造的な相違がある 2 つのタンパク質の間の接続を新たに発見できる可能性がある．このような接続のカタログは，ゲノムの機能的ルックアップテーブルとして使用することができる．この接続カタログのデータベースとして，Broad Institute が提供する接続マップ（Connectivity Map）がある[9, 10]．Connectivity Map では複数の細胞型で試験された，約 5000 の小分子化合物，約 3000 の遺伝子試薬からの 1.5 M 以上の遺伝子発現プロファイルからなるライブラリが作成されている．

■参考文献

1) https://www.ncbi.nlm.nih.gov/geo/
2) T. Barrett, S.E. Wilhite, P. Ledoux, C. Evangelista, I.F. Kim, M. Tomashevsky, K.A. Marshall, K.H. Phillippy, P.M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C.L. Robertson, N. Serova, S. Davis, and A. Sobolev :“ NCBI GEO: archive for functional genomics data sets–update, ”Nucleic Acids Res., 41, pp.D991-D995, 2013.
3) https://cancergenome.nih.gov/
4) Cancer Genome Atlas Research Network :“ Comprehensive genomic characterization defines human glioblastoma genes and core pathways, ” Nature, 455(7216), pp.1061-1068, 2008.
5) https://www.ebi.ac.uk/arrayexpress/
6) N. Kolesnikov, E. Hastings, M. Keays M, O. Melnichuk, Y.A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma :“ ArrayExpress update–simplifying data submissions, ”Nucleic Acids Res., 43, pp.D1113-D1116, 2014.
7) https://www.encodeproject.org/

8) C.A. Davis, B.C. Hitz, C.A. Sloan, E.T. Chan, J.M. Davidson, I. Gabdank, J.A. Hilton, K. Jain, U.K. Baymuradov, A.K. Narayanan, K.C. Onate, K. Graham, S.R. Miyasato, T.R. Dreszer, J.S. Strattan, O. Jolanki, F.Y. Tanaka, and J.M. Cherry :“ The Encyclopedia of DNA elements (ENCODE): data portal update, ” Nucleic Acids Res., 46(D1), pp.D794-D801, 2018.
9) https://www.broadinstitute.org/connectivity-map-cmap
10) J. Lamb :“ The Connectivity Map: a new tool for biomedical research, ” Nature Reviews Cancer, 7, pp.54-60, 2007.

■S2 群 - 6 編 - 4 章

4-8 文献情報

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

4-8-1 PubMed

PubMed[1, 2]は，米国の国立衛生研究所（NIH）にある米国国立医学図書館（NLM）の国立バイオテクノロジー情報センター（NCBI）によって開発・維持されている無料の文献検索サービスである．Pubmed では，MEDLINE，ライフサイエンス関連の学会誌，オンラインで参照できる書籍からの生物医学文献への参照情報を取り扱っており，その数は 2600 万件を超えている（2017 年 9 月現在)．検索の対象分野は，生物医学と健康の分野であり，主にライフサイエンス，行動科学，化学，バイオエンジニアリングが検索対象となる．PubMed は，また，関連するほかのウェブサイトへのアクセスを提供し，ほかの NCBI 分子生物学データベースへのリンクも提供する．

4-8-2 Google Scholar

Google Scholar[3]は，学術文献を幅広く検索するための Google が提供するサービスである．検索対象は学術出版社，専門家団体，オンラインリポジトリー，大学，その他のウェブサイトであり，記事，論文，書籍，要約など，様々な分野である．Google の持つ強力な検索機能を利用しているため，学術文献を検索できるだけでなく，関連する研究や，引用，著者，出版物に関する情報も調べることができ，関連する研究分野の最新の動向を把握することも容易となっている．

4-8-3 Microsoft Academic

Microsoft Academic [4]は，学術文献を幅広く検索するための Microsoft が提供するサービスである．Microsoft が提供するセマンティック検索技術により，1 億 2000 万件を超える検索対象から，検索内容に関して，最新かつ高い関連性を持つ情報を検索することができる．

■参考文献

1) https://www.ncbi.nlm.nih.gov/pubmed/
2) NCBI Resource Coordinators :“ Database resources of the National Center for Biotechnology Information, ” Nucleic Acids Res., 44, pp.D7-D19, 2016.
3) https://scholar.google.co.jp
4) http://academic.research.microsoft.com/

■S2 群 - 6 編 - 4 章

4-9 GWAS（Genome Wide Association Study）

（執筆者：稲岡秀檢）[2018 年 2 月 受領]

遺伝学において，ゲノムワイド関連研究（GWAS）は，異なる個体におけるゲノムワイドな一連の遺伝子変異情報を利用する解析手法である．遺伝子に変異体が生じると，その遺伝子が関わる形質が変化する可能性がある．そこから形質が変化した群と，形質が変化していない対照群に分けて，両者の間で異なる遺伝子を探せば，それが原因遺伝子である可能性がある．しかし，単体の一塩基多型（SNP）を調べるだけでは，疾患のような形質の変化を説明することは困難である．GWAS とは，ゲノム全体をカバーするような多数の SNP のデータベースを利用し，形質と SNP の頻度との関連を統計的に検査することで求める方法のことである．

症例対照ゲノムワイド関連研究により，疾患関連変異体を見出すためには、少なくとも数千の症例と数千の対照群が十分な検出力のために必要であり，ヒトゲノムを十分にカバーするためには数十万またはそれ以上の SNP の情報が必要である[1)]．それゆえ，GWAS ではデータの規模がかなり大きい．その結果，GWAS では計算上効率的でかつ利用可能なデータを有効に活用できる新しい分析方法が必要となる．

GWAS の能力を向上させる一つの方法は，単一のマーカ SNP と形質の関連検査だけでなく，ハプロタイプの組合せを推測して，ハプロタイプベースの関連検査を行うことである．ハプロタイプとは，各遺伝子座位にある対立遺伝子のいずれか一方の組合せをいう．例えば，ある遺伝子座位（座位 1）の対立遺伝子が A または a，別の遺伝子座位（座位 2）の対立遺伝子が B または b である場合を考える．二倍体生物の場合，座位 1 として AA，Aa，aa の 3 つのパターンが考えられる．同様に座位 2 として BB，Bb，bb の 3 つのパターンが考えられる．

上記の 3×3 の組み合わせた遺伝子型のうち，ヘテロ同士の組合せ以外，つまり Aa，Bb 以外の座位 1 と座位 2 の組合せとして得られる遺伝子型については，AB-AB，AB-Ab，AB-aB，ab-ab，ab-Ab，ab-aB，aB-aB，Ab-Ab は一意に決定できる．ただし，AB-Ab と Ab-AB のように順番が違うものも同じ遺伝子型とする．

しかし，Aa，Bb の組合せの場合，AB-ab あるいは Ab-aB のいずれかにになり，一意に決定できない．

ハプロタイプの組合せを統計的に推定する方法として，隣接する変異体間の相関である連鎖不平衡を用いる方法がある．推定には隠れマルコフモデルが使用されることが多い．

もう一つの方法は，欠損したデータのインピュテーション（補完）を使用して，既知の変異体（ただし，まだ遺伝子型が知られていない）の遺伝子型を推測することである．現在，インピュテーションに使用される変異体としては，HapMap プロジェクトで遺伝子型が決定された SNP のデータベースが用いられる．

以下に，ハプロタイプ推定で使用されるツールと欠損データのインピュテーションに使用されるツールを紹介する．

4-9-1 ハプロタイプ推定ツール

GWAS において，ハプロタイプ推定は，ハプロタイプベース関連研究を行うために使用さ

れる．可能性のあるハプロタイプの組合せは膨大となる．そのため，GWAS データの計算上の負担を軽減するためには，各個人のハプロタイプの組合せの最良推定値だけで解析を行う必要がある．

ハプロタイプ推定に用いられるツールの多くは Li と Stephens が提唱した Product of Approximate Conditionals（PAC）モデル[2]に基づいている．直接観測できるのは遺伝子型のみであり，真のハプロタイプは推定することしかできない．このモデルでは，データベースから得られる参照ハプロタイプの幾つかのサブセットを利用する．それぞれの参照ハプロタイプは，各マーカ SNP における隠れマルコフモデルの状態を表していると考えられる．

FastPHASE[3]は，参照ハプロタイプの代わりに一定数のハプロタイプクラスタを有するフレームワークを使用する．ハプロタイプクラスタの定義，組換え及び突然変異率の定義を含むモデルパラメータは，期待値最大化アルゴリズムを使用して推定される．

Mach[4]も Li と Stephens フレームワークに基づいているツールである．推定において期待値最大化法を繰り返し，推定されたハプロタイプの組合せを参照ハプロタイプとして使用する．参照ハプロタイプから，個体を一つずつ除去して，推定を繰り返し結果を更新していく．

PHASE[3]は，小さなデータセットで優れた結果を達成したため，ハプロタイプ推定の分野で広く用いられている．PHASE は，Bayes アプローチを採用し，マルコフ連鎖モンテカルロアルゴリズムを使用してモデルのパラメータフィッティングを行う．ただし，マルコフ連鎖モンテカルロアルゴリズムを使用しているため，長い計算時間が必要となる．

4-9-2 欠落データのインピュテーションツール

ハプロタイプ推定法は，ハプロタイプの組合せを推論する過程で欠損データのインピュテーションを行っている．Impute[5]はハプロタイプが推定された参照ハプロタイプを使用してモデルを構築する．組換え率と突然変異率はインピュテーションを行うユーザが指定する．そのため，インピュテーションに反復的なモデル構築アプローチを必要としないが，モデルパラメータの誤指定に敏感である．また，インピュテーションが行われる個体に関する情報も利用しないという特徴がある．

Bim-Bam[6]は，fastPHASE を使用してインピュテーションを実行する．欠損データは複数回インピュテーションされる．インピュテーションされた値は関連検査をするためのベイジアン回帰法で利用される．

4-9-3 GWAS Catalog

現在までに行われた GWAS 研究より得られた，形質変化に関係があると推定された SNP に関する情報が GWAS Catalog[7, 8]にて公開されている．

論文として公開された研究から得られた約 10 万の SNP と，形質との関連があると考えられる p 値が 1.0×10^{-5} 以下の SNP について，文献情報を元に手作業で分類されたデータベースである．形質に関連すると考えられる SNP を染色体上の位置情報を使って位置をマッピングし，SNP の存在位置を視覚的に表現するツールも提供されている．

■参考文献

1) S.R. Browning：“ Missing data imputation and haplotype phase inference for genome-wide association studies, ” Hum Genet., 124(5), pp.439-450, 2008.
2) N. Li, M. Stephens：“ Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, ” Genetics, 165(4), pp.2213-2233, 2003.
3) P. Scheet, and M. Stephens：“ A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase, ” Am J Hum Genet., 78(4), pp.629-644, 2006.
4) A. Vaez, R. Jansen, B.P. Prins, J.J. Hottenga, E.J. de Geus, D.I. Boomsma, B.W. Penninx, I.M. Nolte, H. Snieder, B.Z. Alizadeh：“ In Silico Post Genome-Wide Association Studies Analysis of C-Reactive Protein Loci Suggests an Important Role for Interferons, ” Circ Cardiovasc Genet., 8(3), pp.487-497, 2015.
5) Z. Yu, D. Schaid：“ Methods to impute missing genotypes for population data, ” Hum Genet., 122(5), pp.495-504, 2007.
6) L.J. Scott, K.L. Mohlke, L.L. Bonnycastle, C.J. Willer, Y. Li, W.L. Duren, M.R. Erdos, H.M. Stringham, P.S. Chines, A.U. Jackson, L. Prokunina-Olsson, C.J. Ding, A.J. Swift, N. Narisu, T. Hu, R. Pruim, R. Xiao, X.Y. Li, K.N. Conneely, N.L. Riebow, A.G. Sprau, M. Tong, P.P. White, K.N. Hetrick, M.W. Barnhart, C.W. Bark, J.L. Goldstein, L. Watkins, F. Xiang, J. Saramies, T.A. Buchanan, R.M. Watanabe, T.T. Valle, L. Kinnunen, G.R. Abecasis, E.W. Pugh, K.F. Doheny, R.N. Bergman, J. Tuomilehto, F.S. Collins, and M. Boehnke：“ A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants, ” Science, 316(5829), pp.1341-1345, 2007.
7) https://www.ebi.ac.uk/gwas/
8) J. MacArthur, E. Bowler, M. Cerezo, L. Gil, P. Hall, E. Hastings, H. Junkins, A. McMahon, A. Milano, J. Morales, Z. Pendlington, D. Welter, T. Burdett, L. Hindorff, P. Flicek, F. Cunningham, and H. Parkinson：“ The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog), ” Nucleic Acids Res., 45, pp.D896-D901, 2017.

■S2 群 - 6 編 - 4 章

4-10 各種ツール

（執筆者：稲岡秀檢）［2018 年 2 月 受領］

4-10-1 データ表示ツール

例えば，特定の遺伝子について解析した結果を表示するときに，その遺伝子がどの染色体上のどの位置にあるのか，近傍にはどのような遺伝子があるのかが，同時にグラフィカルに表示されると解析結果に対する理解がより深まる．また，解析した遺伝子に関して，過去にどのような研究が行われ，どのような文献情報があるのかなどについても同時に表示されると，新たな視点からの研究の推進のための手助けともなる．こういった情報開示が行われている例として NCBI の Genome Data Viewer [1)]がある．ここではヒトを含む多くの生物種の遺伝子情報が検索可能であり，上記のような情報をグラフィカルに閲覧することが可能である．

このようなデータの視覚化をサポートとするツールの一つとして GBrowse [2)]がある．GBrowse はゲノム情報を表示・操作するためのデータベースと対話的な Web ページを組み合わせたものである．

その機能の特徴としては，俯瞰図と詳細の図を同時に表示する，スクロールやズーム，URL やアノテーションの付加，検索，複数言語のサポート，カスタマイズ可能といったものがある．

そのほかに，データの視覚化が有効な例としてはタンパク質のデータ解析がある．タンパク質の機能はタンパク質の三次元構造に深く関与している．そのため，タンパク質の機能解析においては，タンパク質の構造を三次元で可視化することは有効である．このような可視化ツールの一つとして，Swiss-Pdb Viewer [3, 4)]がある．

リガンド-標的タンパク質の複合体の構造決定は，リガンド-標的タンパク質標の相補性を理解する非常に重要な情報と考えらる．複合体の構造が解明されていて利用可能な 3D 構造が存在する場合，構造を詳細に観察することにより，構造に関する深い理解を得ることができるが，複合体の研究科にとってもその観測を迅速に行うことは難しい．専門家ではなく，医薬品化学などのほかの分野のユーザは，最初に 3D ビジュアライゼーションソフトウェアを理解することが必要になる．このことは，貴重な 3D 構造情報を共有することに対する障壁である．このような問題を解決するために作成された 3D ビジュアライゼーションソフトウェアとして LigPlot+ [5)]がある．

LigPlot+では Protein Data Bank（PDB）のすべての構造モデルのリガンド-タンパク質相互作用を表示することができ，更に描画された図をインタラクティブに編集することが可能である．この結果，複合体の特定の部位を回転させたり，ある特定の結合について色を変更したり，テキストラベルを作成することができる．

また，複数のリガンド複合体を表示することが可能で，異なるタンパク質に結合する単一のリガンドや，同じタンパク質に結合する複数のリガンドなどを容易に検討できるため，一般的な研究だけでなく，薬理ゲノミクスなどの幅広い事項に適用できる．

4-10-2 プログラミング言語

バイオインフォマティクスにおいてデータ解析で使用されるプログラミング言語は多数あるが，そのなかでも使用頻度が高いものについて概説する．

（1）BioPerl

BioPerl [6]とは，Perl [7]という文字列処理に適したインタープリタ型のプログラミング言語をベースとして，バイオインフォマティクス用の様々な拡張ライブラリを実装したプログラミング処理系である．

BioPerl では，各種データベースの塩基配列，ペプチド配列データへアクセスや，ほかのデータベースで利用されている形式への変換，塩基配列に対する操作や，類似の配列の検索などバイオインフォマティクスで求められる各種ライブラリが豊富に実装されている．

4-10-3 BioRuby

BioRuby [8]とは，Ruby [9]というオブジェクト指向のインタープリタ型のプログラミング言語をベースとして，バイオインフォマティクス用の様々な拡張ライブラリを実装したプログラミング処理系である．BioPerl と同様に DNA やタンパク質の配列解析や配列アライメント，各種データベースへのアクセスなど，バイオインフォマティクスで求められる各種ライブラリが豊富に実装されている．

4-10-4 BioPython

BioPython [10]とは，Python [11]というオブジェクト指向のインタープリタ型のプログラミング言語をベースとして，バイオインフォマティクス用の様々な拡張ライブラリを実装したプログラミング処理系である．BioPerl と同様に DNA やタンパク質の配列解析や配列アライメント，各種データベースへのアクセスなど，バイオインフォマティクスで求められる各種ライブラリが豊富に実装されている．Python には NumPy [12]や SciPy [13]という科学技術計算に特化したライブラリも提供されており，多次元配列処理など高度な科学技術計算も可能となっている．

4-10-5 Bioconductor

Bioconductor [14]とは，R[?]という統計処理に適したインタープリタ型のプログラミング言語をベースとして，バイオインフォマティクス用の様々な拡張ライブラリを実装したプログラミング処理系である．BioPerl と同様に，DNA やタンパク質の配列解析や配列アライメント，各種データベースへのアクセスなど，バイオインフォマティクスで求められる各種ライブラリが豊富に実装されている．また，R は豊富なグラフィック処理ライブラリが実装されており，処理結果の 2D，3D 表示などデータ表示ツールとしても利用可能である．

■参考文献

1) https://www.ncbi.nlm.nih.gov/genome/gdv/
2) http://gmod.org/wiki/Main_Page
3) http://spdbv.vital-it.ch/
4) M.U. Johansson, V. Zoete, O. Michielin, and N. Guex :“ Defining and searching for structural motifs using DeepView/Swiss-PdbViewer, ” BMC Bioinformatics, 13:173, 2012.
5) R.A. Laskowski and M.B. Swindells :“ LigPlot+: multiple ligand-protein interaction diagrams for drug discovery, ” J Chem Inf Model., 51(10), pp.2778-2786, 2011.
6) http://bioperl.org/

7) https://www.perl.org/
8) http://bioruby.org/
9) https://www.ruby-lang.org/
10) http://biopython.org/
11) https://www.python.org/
12) http://www.numpy.org/
13) https://www.scipy.org/
14) https://www.bioconductor.org/
15) https://www.r-project.org/