平成２１年度科学技術試験研究委託事業
「生体分子の熱力学データと構造データの統合」

２１年度　委託業務研究成果報告書

平成２２年３月

国立大学法人九州工業大学　大学院情報工学研究院　教授　皿井明倫

本報告書は、文部科学省の科学技術試験研究委託事業による委託業務として、国立大学法人九州工業大学が実施した平成２１年度「生体分子の熱力学データと構造データの統合」の成果を取りまとめたものです。

目次

１．委託業務の目的・・・・・・・・・・・・・・・・・・・・・・・・・・・４
２．平成２１年度（報告年度）の実施内容　・・・・・・・・・・・・・・・・４
　２．１ 実施計画・・・・・・・・・・・・・・・・・・・・・・・・・・・４
　２．２ 実施内容(成果)・・・・・・・・・・・・・・・・・・・・・・・・５
　２．３ 成果の外部への発表・・・・・・・・・・・・・・・・・・・・・・７
　２．４ 活動（運営委員会等の活動等)・・・・・・・・・・・・・・・・・ ８
　２．５ 実施体制・・・・・・・・・・・・・・・・・・・・・・・・・・・９
別紙参考資料１・・・・・・・・・・・・・・・・・・・・・・・・・・１０
別紙参考資料２・・・・・・・・・・・・・・・・・・・・・・・・・・１７
別紙参考資料３・・・・・・・・・・・・・・・・・・・・・・・・・・１８
別紙参考資料４・・・・・・・・・・・・・・・・・・・・・・・・・・１９
別紙参考資料５・・・・・・・・・・・・・・・・・・・・・・・・・・２１
整備実績一覧・・・・・・・・・・・・・・・・・・・・・・・・・・・２２

１．委託業務の目的

中核機関である情報・システム研究機構では、ライフサイエンスやバイオ産業に従事する研究者や技術者がいわゆるゲノムプロジェクト・ポストゲノムプロジェクトの成果や多様な DB や解析ツールをストレスなく利用してより高度な研究開発が効率よくできる環境（統合 DB）を実現することを目的とする。このため、参画機関と共同で、戦略立案・実行評価、統合データベース開発、および、統合データベース支援を行うこととしている。

本研究では、情報・システム研究機構が進める統合化を補完するため、蛋白質の安定性や相互作用の網羅的な熱力学データを構造データと統合する。これにより、生体分子の機能に関する研究を促進する。また、構造データベースを構築する PDBj とも連携して、XML などのデータ交換フォーマットの整備、オントロジーなどの統合化技術の開発を行う。さらに、情報・システム研究機構による統合検索との連携を可能にするために、情報・システム研究機構と連携して開発を進める

２．平成２１年度（報告年度）の実施内容

２．１ 実施計画

①蛋白質と変異体の熱力学データベースの構築と統合

平成２１年度に収集する蛋白質およびその変異体の構造安定性に関する熱力学データ（約1,500件を予定）について、熱力学データと構造データを対応づけるクロスレファレンステーブルを作成する。また、情報・システム研究機構による統合検索と連携するため、データの提供、インデックス作成などを行う。一方、情報・システム研究機構と連携して、テキストマイニングによるデータの自動収集や自動抽出法の開発を進める。

②蛋白質・核酸相互作用の熱力学データベースの構築と統合

平成２１年度に収集する蛋白質と核酸の相互作用の定量的な熱力学実験データ（約1,300件を予定）について、蛋白質・核酸複合体の構造データがあるものについて熱力学データと構造データを対応づけるクロスレファレンステーブルを作成する。また、情報・システム研究機構による統合検索と連携するため、データの提供、インデックス作成などを行う。一方、情報・システム研究機構と連携して、テキストマイニングによるデータの自動収集や自動抽出法の開発を進める。

③蛋白質･蛋白質相互作用データの生成と統合

蛋白質･蛋白質相互作用データを格納するデータベースのプロトタイプ（ＤＢスキーマおよび関係データベースシステムへの実装）を作成する。

④XMLデータフォーマットやオントロジーなどの統合化技術の開発

平成２０年度に引き続きオントロジーの整備を進める。熱力学データのXMLフォーマットを作成し公開する。また、XMLと他のフォーマットの変換を行う

ためのプログラムを作成する。これらの開発にあたっては、構造データの我が国の代表機関としてすでに活動しているPDBjと連携して進める。

２．２ 実施内容(成果)

① 蛋白質と変異体の熱力学データベースの構築と統合

平成２１年度に発生した蛋白質およびその変異体の構造安定性に関する熱力学データ約 1,000 件について、熱力学データと構造データを対応づけるクロスレファレンステーブルを作成した。すなわち、熱力学データベースに記載された PDB の構造データ（PDBcode）およびそれと 100%同じ配列の PDBcode と対応するすべての熱力学データをリストしたテーブルと、配列の類似（95%以上の類似度）するすべての構造と対応するすべての熱力学データをリストしたテーブルを作成した。

蛋白質と変異体の熱力学データベースの検索画面とデータベースの内容については、それぞれ別紙参考資料１-(1)、１-(3)を参照。クロスレファレンステーブルについては、別紙参考資料１-(4)を参照。なお、蛋白質と変異体の熱力学データベース、ProTherm、のアクセス統計データを別紙参考資料１-(6)に示した。アクセスは、月当たり２千～４千件程度である。利用者層としては、企業からのアクセスが過半数を占めている。国別では、国内、アメリカ、欧州などからのアクセスが多い。本データベースはすでに公開から１２年がたち、約 25,000 件のデータを保有しているので、我々が開発したデータベースの中では最もアクセスが多い。利用者は専門あるいは関連分野の研究者であるので、アクセスのほとんどは検索を行ってデータのページを参照している。このデータベースのデータは世界中の研究者に利用され、その成果が発表された論文はすでに 200 件以上になっている。論文のリストは以下の URL を参照。

http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/ProTherm_References.htm

本研究が統合の対象としているデータベースの構築にあたっては、熱力学データの含まれている文献を収集し、論文を研究者が読んでデータの抽出を行っている。その後データの入力から照合・チェック、データベースへの登録までをほとんど手動で行っている。特に、研究者が論文からデータを抽出する部分が最も手間がかかる作業となっている。そこで、テキストマイニングの手法などを取り入れて、文献の自動収集や文献からのデータの自動抽出を行い、データベース構築の省力化を計ろうとしている。情報・システム研究機構では、そのような目的のためにテキストマイニングツール、TogoDoc や Wired-Marker を開発している。本研究では、これらのツールが我々の作業に応用できるかどうかの評価を行った。TogoDoc については類似文献検索、Wired-Marker ではテキストからの情報の自動抽出について我々のデータベース構築の有効性の観点から評価を行った。Wired-Marker については、HTML 形式での論文の表の任意の行と列からのデータの抽出、特殊文字の扱い、図からのデータ抽出、XML からの必要データの抽出、PDF の文献からのデータ抽出、などの方法について検討を行った。また TogoDoc については、まず、これまでに我々が収集したデータの記載された文献、蛋白質名、キーワードのリストや、マーキングした論文のサンプルなどをセンター側に提供した。これらの情報をもとに、TogoDoc において PubMed の related articles の機能を用いて類似文献を検索し、データを含む文

献がどれだけヒットするかの検証を行った。また、TogoDoc 独自の学習機能を備えた検索ツールを評価するため、これまでの正例の文献とともにデータを含まない負例の文献を抽出し提供した。今後これらのサンプルを用いて、学習のチューニングや検証を行う予定である。

② 蛋白質・核酸相互作用の熱力学データベースの構築と統合

平成２１年度に今年度発生した蛋白質と核酸の相互作用の定量的な熱力学実験データ約 1,300 件について、蛋白質・核酸複合体の構造データがあるものについて熱力学データと構造データを対応づけるクロスレファレンステーブルを作成した。すなわち、熱力学データベースに記載された PDB の構造データ（PDBcode）およびそれと 100%同じ配列の PDBcode と対応するすべての熱力学データをリストしたテーブルと、配列の類似（95%以上の類似度）するすべての構造と対応するすべての熱力学データをリストしたテーブルを作成した。

蛋白質・核酸相互作用の熱力学データベースの検索画面とデータベースの内容については、それぞれ別紙参考資料１-(1)、１-(3)を参照。クロスレファレンステーブルについては、別紙参考資料１-(5)を参照。蛋白質・核酸相互作用の熱力学データベース、ProNIT、のアクセス統計データは別紙参考資料１-(6)に示した。アクセスは、月当たり２千～３千件程度である。利用者層としては、やはり企業からのアクセスが過半数を占める。国別では、国内、アメリカ、アジア、欧州などからのアクセスが多い。利用者は専門あるいは関連分野の研究者であるので、アクセスのほとんどは検索を行ってデータのページを参照している。このデータベースを利用してその成果が発表された論文のリストは以下の URL を参照。

http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit_ref.html

テキストマイニングについては、蛋白質と変異体の熱力学データベースの部分と同様、情報・システム研究機構と連携して、テキストマイニングツール、TogoDoc、Wired-Marker の評価を行った。詳細は、前節「蛋白質と変異体の熱力学データベースの構築と統合」の部分を参照。

③ 蛋白質・蛋白質相互作用データの生成と統合

蛋白質･蛋白質相互作用データを格納するデータベースのプロトタイプを作成した。データベーススキーマを別紙参考資料２に示す。

④ XML データフォーマットやオントロジーなどの統合化技術の開発

熱力学データと構造データの統合を効率的にすすめるため、熱力学データに関するオントロジーについて引き続き調査を行った。生命情報に関してはすでに多くのドメインでオントロジーが整備されつつあるが、熱力学データに関するオントロジーはまだ整備されていないので、我々が構築している熱力学データベースについて Controlled Vocabulary の作成を継続して行った（別紙参考資料３を参照）。具体的には、本年度に行ったスキーマの改良などに対応して Controlled Vocabulary の改良を行った。なお、オントロジー整備にあたっては、構造データベースの代表機関である PDBj や海外の関係機関と意見交換を行っている。

これまではフラットデータのみを公開してきたが、平成２０年度にフラットデータを XML フォーマットのプロトタイプに変換するプログラムを作成し試験的に公開した。平成２１年度は、XML フォーマットを作成するプログラムを完成させ、XML フォーマットのデータを９月に完全公開した。さらに平成２１年度は、XML フォーマットデータをテキストフォーマットに変換するプログラムも作成し、テキストフォーマットデータも XML フォーマットと同時に公開した。XML データのサンプルは別紙参考資料４-(1)を参照。変換のステップは別紙参考資料４-(2)を参照。

なお、本プロジェクトに関する最新情報は、プロジェクト専用の Web ページ（別紙参考資料５を参照）を通して公開している。

２．３ 成果の外部への発表

論文寄稿

業務コード	実施年度	和誌/洋誌	論文タイトル	発表者名	発表誌名	巻	号	ページ	掲載年月	メモ
1〇	2009	洋	Thermodynamic Database for Proteins: Features and Applications	M. M. Gromiha and A. Sarai	Methods Mol. Biol.	609		97-112	2010	

講演

業務コード	実施年度	国内/国際	講演タイトル	発表者名	講演会名	発表年月日	メモ
1◎	2009	国内	生体分子の熱力学データと構造データの統合	皿井明倫、Shaji Kumar	文科省統合データベースプロジェクト シンポジウム「データベースが拓くこれからのライフサイエンス」	2009 年 6 月 12 日	

プレス発表

業務コード	実施年度	発表タイトル	掲載新聞名	掲載日

２．４ 活動（運営委員会等の活動等）

中核機関との会合の履歴

件名：テキストマイニング、横断検索などについて
日時：2009 年 5 月 12 日（火） 16:00～17:00
場所：ライフサイエンス統合データベースセンター・九州工業大学間の TV 会議
参加者：九工大：皿井、Kumar、統合 DB センター：西川、畠中、平井

件名：研究運営委員会作業部会分科会
日時：2009 年 5 月 29 日（金）10:00～12:00
場所：ライフサイエンス統合データベースセンター

件名：テキストマイニングについて
日時：2009 年 5 月 29 日（金）15:30～16:45
場所：ライフサイエンス統合データベースセンター
参加者：九工大：皿井、Kumar、統合 DB センター：西川、畠中、山本、平井

件名：テキストマイニング、データの利用許諾について
日時：2010 年 2 月 2 日（火）12:00～13:00
場所：ライフサイエンス統合データベースセンター
参加者：九工大：皿井、Kumar、統合 DB センター：畠中、山本、平井

件名：研究運営委員会作業部会分科会
日時：2010 年 2 月 2 日（火）13:30～16:30
場所：ライフサイエンス統合データベースセンター

件名：テキストマイニングについて
日時：2010 年 3 月 8 日（月）13:00～14:00
場所：ライフサイエンス統合データベースセンター
参加者：九工大：皿井、統合 DB センター：畠中、山本、平井

２．５ 実施体制

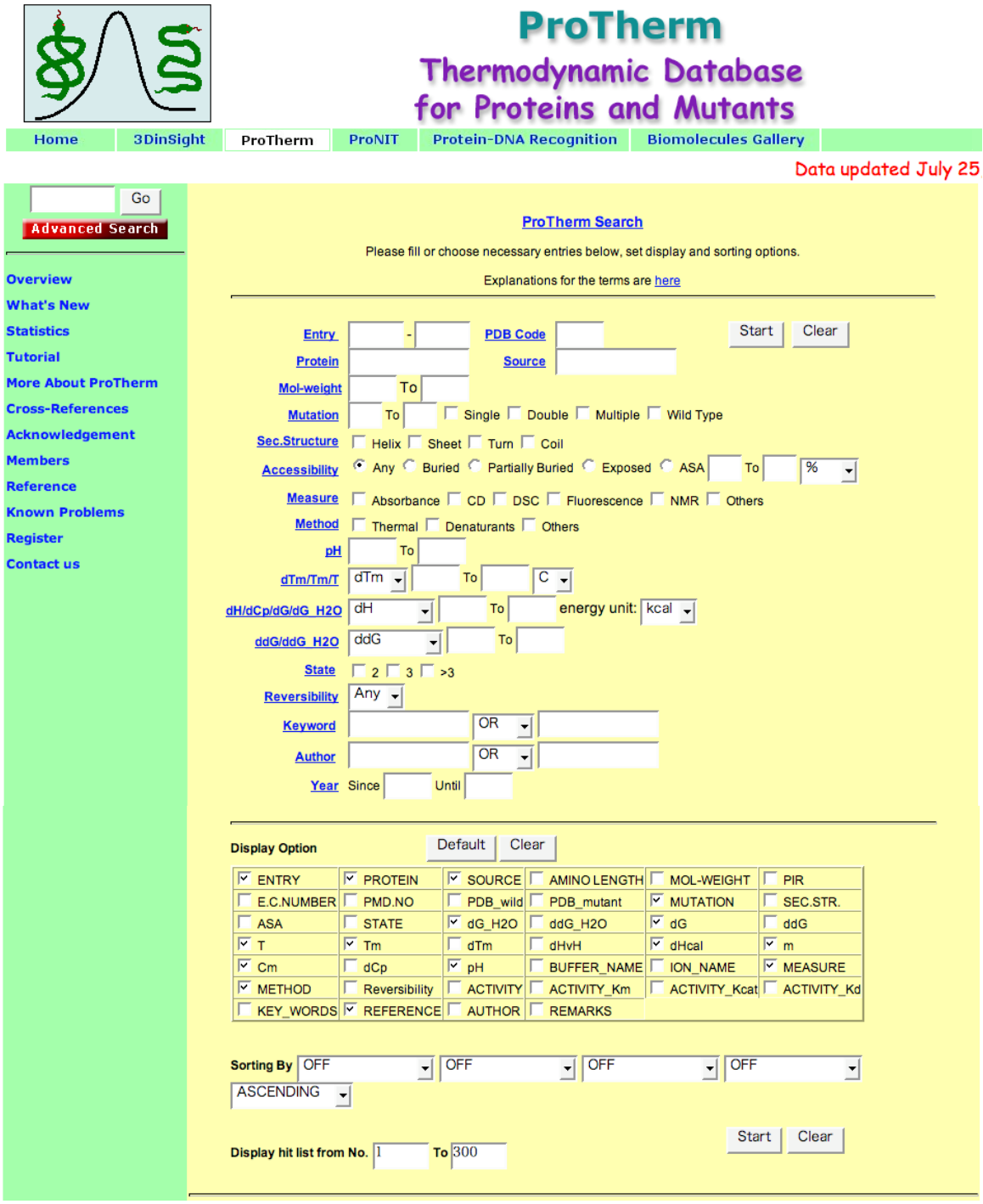
研 究 項 目	担当機関等	研究担当者
1. 蛋白質と変異体の熱力学データベースの構築と統合	九州工業大学情報工学部	◎○皿井 明倫
2. 蛋白質・核酸相互作用の熱力学データベースの構築と統合	九州工業大学情報工学部	○Shaji Kumar
3. 蛋白質・蛋白質相互作用データの生成と統合	九州工業大学情報工学部	○末田 慎二
4. XML データフォーマットやオントロジーなどの統合化技術の開発	九州工業大学情報工学部	○藤井 聡

注1. ◎:課題代表者、○:サブテーマ代表者

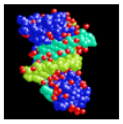
注2. 本業務に携わっている方は、全て記入。

別紙参考資料１

（１）蛋白質熱力学データベース ProTherm の検索画面



（２）蛋白質・核酸相互作用熱力学データベース ProNIT の検索画面



ProNIT
Thermodynamic Database for Protein-Nucleic Acid Interactions

Home | 3DinSight | ProTherm | ProNIT | Protein-DNA Recognition | Biomolecules Gallery

Last Update: 28-Aug.-2007, NEW RELEASE: ProNIT 2.0

Quick Search

- ProNIT Home
- What's New
- About ProNIT
- Release Notes
- Statistics
- Tutorial
- Members
- Reference
- Contact us
- Copyright
- Acknowledgement

Welcome to ProNIT Database

Advanced Search

Please fill or choose necessary entries below, set display and sorting options.

ProNIT Entry - Search Reset Entry List
Protein Name Protein List
Protein Source Source List
PDB_Code PDB List
Protein Sequence
Mutation To Single Double Multiple Wild
Sec.Str Helix Sheet Turn Coil
ASA_Free To %
Nucleic Acid Name
Nucleic Acid Source
Nucleic Acid Type dsDNA ssDNA RNA
NDB_Code NDB List
Nucleic Acid Sequence
Method 1.Calorimetry 2.Footprint 3.Filter binding 4.Fluorescence
T To °C
pH To
Kd x 10 To x 10 M
dG To kcal/mol
dH To kcal/mol
dCp To kcal/mol/K
Author AND Author List
Year To Reference List
Keywords AND

Display Options

Protein Name	Protein Source	Biological_unit	Fragment	E.C.Number
PIR_No.	SWISSPROT_NO	PDB_Free	PDB_Complex	NDB_Complex
Mutation protein	ASA_Free	ASA_Complex	ProTherm_No.	Sec_Str
Nucleic Acid Name	Nucleic Acid Source	Nucleic Acid Type	GenBank_No.	Complex_DB_No
Ligand	T	pH	Buffer_Name	Buffer_Conc
Additives	Ion Name	Ion_Conc	Method	Kd_Wild
Kd_Mutant	Ka_Wild	Ka_Mutant	dG_Wild	dG_Mutant
dH_Wild	dH_Mutant	dCp_Wild	dCp_Mutant	Stoichiometry
Activity_Km	Activity_Kcat	Author	Reference	

Sorting Options

Priority 1	Priority 2	Priority 3	Priority 4	Order
OFF	OFF	OFF	OFF	ASCENDING

Entries per page : 50 Search Clear

（３）熱力学データベースの内容

①蛋白質熱力学データベース ProTherm に含まれる主な内容は以下のようである。蛋白質情報：名前、由来種、対応する配列や構造の ID、天然状態における集合数など。変異情報：変異アミノ酸とその位置、２次構造と Accessible Surface Area (ASA)など。実験情報：測定方法や、温度、pH、バッファー、イオン、蛋白質濃度などの実験条件。熱力学データ：熱変性の場合、変性の自由エネルギー変化（ΔG）、エンタルピー変化（ΔH）、熱容量変化（ΔC_p）、変性温度（T_m）、変性の可逆性、変性剤変性の場合、変性剤濃度ゼロに外挿した変性自由エネルギー変化（ΔG^{H2O}）、変性曲線の傾き（m）と変性中点の変性剤濃度（C_m）など。その他の情報：酵素活性値（K_m、k_{cat}）、解離定数（K_d）、転移の状態数。文献情報：ジャーナル名、著者名、出版年、キーワード、リマークなど。

②蛋白質・核酸相互作用熱力学データベース ProNIT に含まれる主な内容は以下のようである。蛋白質情報: 名前、由来種、対応する配列や構造の ID など。アミノ酸変異情報：変異アミノ酸とその位置、２次構造と ASA など。核酸情報：名前、由来種、対応する配列や構造などの ID。塩基変異情報：変異塩基とその位置。複合体情報：複合体構造の ID、複合体形成に伴う構造変化などの記述。実験情報: 測定方法や、温度、pH、バッファー、イオン、蛋白質濃度などの実験条件。熱力学データ：解離定数（K_d）、結合の自由エネルギー変化（ΔG）、エンタルピー変化（ΔH）、熱容量変化（ΔC_p）、結合の stoichiometry。その他の情報：酵素活性値（K_m、k_{cat}）。文献情報：ジャーナル名、著者名、出版年、キーワード、リマークなど。

（４）クロスレファレンステーブル。ProTherm データベースに含まれる構造データの PDBcode と 100%同じ配列に対応する蛋白質の熱力学データの対応表の一部。

Cross-Reference to PDB

PDB_ID	ProTherm_EntryNo
1BAL	18200, 18201, 18202, 18203, 18204, 18205, 18206, 18207, 18208, 18209
2TCT	6413, 6414
1IKL	23670, 23671, 23672, 23673, 23674, 23675
1PUC	10219, 10220, 10221, 10222, 10223, 10224, 10225, 10226, 10227, 10228, 10229, 10230, 10231, 10232, 10233, 10234, 10235
1TIO	7485, 7486, 7487, 7488, 7489, 9882, 16833
1IKM	23670, 23671, 23672, 23673, 23674, 23675
1BAV	15105, 15106
1TIU	15280, 15281, 17002, 17003, 17013, 17014, 17628, 17629, 17630, 17631, 17632, 17633, 17634
1F6H	248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 328, 329, 330, 331, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2812, 2813, 2814, 2815, 2816, 2817, 2818, 2819, 2820, 2821, 2822, 2823, 2824, 2825, 2826, 2827, 2828, 2829, 4545, 4546, 5899, 5900, 5901, 5902, 5903, 5904, 5905, 5906, 5907, 5908, 5909, 5910, 5911, 5912, 6215, 6216, 6503, 6504, 6505, 6506, 6507, 6508, 6509, 6510, 6511, 6512, 6513, 6514, 6515, 6516, 6517, 6518, 6519, 6520, 6521, 6522, 6523, 6524, 6525, 6526, 6527, 6528, 6529, 6530, 6531, 6532, 6533, 6534, 6535, 6536, 6537, 6538, 6539, 6540, 6541, 6542, 6543, 6544, 6545, 6546, 7408, 7409, 7410, 7411, 7493, 7494, 7495, 7496, 7497, 7498, 7499, 7500, 7501, 7502, 7503, 7504, 7505, 7506, 7507, 7508, 7509, 7510, 7511, 7512, 7513, 7514, 7847, 10124, 10139, 10140, 10141, 10142, 10143, 10144, 10145, 10146, 10147, 10148, 10149, 10150, 10151, 10152, 10153, 10154, 10155, 10156, 10157, 10158, 10159, 10160, 11133, 11134, 11135, 11136, 11137, 11138, 11139, 11140, 11141, 11142, 11143, 11144, 11145, 11146, 11147, 11148, 11149, 11150, 11151, 11152, 11153, 11154, 11155, 11156, 11157, 11158, 11159, 11160, 11233, 12362, 13271, 13272, 13273, 13274, 13275, 13276, 13277, 13278, 13279, 13280, 13281, 13282, 13283, 13284, 13285, 13286, 13287, 13288, 13289, 13290, 13291, 13292, 13293, 13328, 13329, 13330, 13331, 13332, 13333, 13334, 13335, 13336, 13337, 13338, 13339, 13340, 13341, 13342, 14518, 14519, 14520, 14521, 14522,

（５）クロスレファレンステーブル。ProNIT データベースに含まれる構造データの PDBcode と 100%同じ配列に対応する蛋白質・核酸相互作用の熱力学データの対応表の一部。

1A1L	2719, 2720, 2721, 2722, 2723, 2724, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 6303, 6306
1A28	5959, 5963
1A3C	5592, 5593, 5594, 5595, 5596, 5597, 5598, 5599, 5600, 5601, 5602, 5603, 5604, 5605, 5606, 5607, 5608, 5609, 5610, 5611, 5612, 5613, 5614, 5615, 5616, 5617, 5618, 5619, 5620, 5621, 5622, 5623, 5624, 5625, 5626, 5627, 5628, 5629, 5630, 5631, 5632, 5633, 5634
1A41	6117, 6118, 6119, 6120, 6121, 6122, 6123, 6124, 6125, 6126, 6127, 6128, 6129, 6130, 6131, 6132, 6133, 6134, 6135, 6136, 6137, 6138, 6139, 6140, 6141, 6142, 6969, 6970, 6971, 6972, 6973, 6974, 6975, 6976, 6977, 6978, 6979, 6980, 6981, 6982, 6983, 6984, 6985, 6986, 6987, 6988, 6989, 6990, 6991, 6992, 6993, 6994, 6995, 6996, 6997, 6998, 6999, 7000, 7001, 7002, 7003, 7004
1A43	8505, 8509, 8512, 8515, 8517, 8519, 8521, 8523, 8525
1A4T	7665, 7666, 7667, 7668, 7669, 7670, 7671, 7672, 7673, 7674, 7675, 7676, 7677, 7678, 7699, 7700, 7701, 7702, 7731, 7732, 7795, 7796, 7797, 7798, 7799, 7800, 7809, 7810, 7811, 7812, 7813, 7814, 7815, 7816
1A4X	5592, 5593, 5594, 5595, 5596, 5597, 5598, 5599, 5600, 5601, 5602, 5603, 5604, 5605, 5606, 5607, 5608, 5609, 5610, 5611, 5612, 5613, 5614, 5615, 5616, 5617, 5618, 5619, 5620, 5621, 5622, 5623, 5624, 5625, 5626, 5627, 5628, 5629, 5630, 5631, 5632, 5633, 5634
1A73	1227, 1228, 1229, 1230, 1231, 1232, 1233, 1234, 1235, 1236, 1237, 1238, 1239, 3364, 3365
1A74	1227, 1228, 1229, 1230, 1231, 1232, 1233, 1234, 1235, 1236, 1237, 1238, 1239, 3364, 3365
1AAB	3173, 3174, 3175, 3176, 3177, 3178, 3179, 3180, 3181, 3182, 3183, 3184, 3185, 3186, 3187, 3188, 3189, 3190, 3191, 3192, 3193, 3194, 3195, 3196, 3197, 3198, 3199, 3200, 3201, 3202, 3203, 3204, 3242, 3243, 3244, 3245, 3246, 3247, 3248, 3249, 3250, 3251, 3252, 3434, 3435, 3436, 3437, 3438, 3439, 3440, 3441
1AAY	2719, 2720, 2721, 2722, 2723, 2724, 2744, 2745, 2746, 2747, 2748, 2749, 2750, 2751, 2752, 2753, 2754, 2755, 2756, 2757, 2758, 2759, 3289, 3290, 3291, 3292, 3293, 3294, 3295, 3296, 3297, 3298, 3299, 6303, 6306
1AF5	3366, 3367, 3368, 3369, 3370, 3371, 3372
1AHD	5938, 5939, 5944, 5945, 5950, 5951
1AIE	1919, 1920, 1921, 1922, 1923, 1924, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 750, 751, 752, 753, 754
1AIS	1557, 1558, 1559, 1560, 1561, 1562, 1563, 1564, 1565, 1566, 1567, 1568, 1569, 1570, 1571, 1572, 1573, 1574, 1575, 1576, 1577, 1578, 1579, 1580, 1581, 1582, 1583, 1584, 1585, 1586, 1587, 1588, 1589, 1590, 1591, 1592, 1593, 1594, 1595, 1596, 1597, 1598, 1599, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 4404, 4405, 4406, 4407, 4408, 4409, 4410, 4411, 4412, 4413, 4414, 4415, 4416, 4417, 4418, 4419, 4420, 4421, 4422, 4423, 4424, 4425,

（６）2009 年度のアクセス推移：

１）ProTherm の月別アクセス統計

月:	リクエスト総数[注1]:	ページリクエスト数[注2]
Apr 2009:	58353:	2495:
May 2009:	41839:	2882:
Jun 2009:	113423:	3447:
Jul 2009:	57265:	3083:
Aug 2009:	70934:	2609:
Sep 2009:	50415:	3344:
Oct 2009:	110411:	4191:
Nov 2009:	154240:	3683:
Dec 2009:	42835:	3434:
Jan 2010:	50563:	3883:
Feb 2010:	64162:	3536:
Mar 2010:	46219:	4113:
Apr 2010:	31391:	2782:

注1)リクエスト総数:アクセスされた HTML ページおよび画像、アイコンなどの総数;注2)ページリクエスト数:アクセスされたユニーク HTML ページの総数

２）ProTherm のアクセスドメイン:TOP10

1552663:	66.48%:	.com (Commercial)
304986:	11.10%:	[unresolved numerical addresses]
145430:	5.02%:	.net (Networks)
125230:	4.21%:	. jp (Japan)
62221:	3.59%:	. edu (USA Higher Education)
28896:	1.06%:	. it (Italy)
25827:	0.76%:	. be (Belgium)
36541:	0.75%:	. tw (Taiwan)
19779:	0.74%:	. ie (Ireland)
28396:	0.73%:	.in (India)

３）ProTherm のリクエストページ：TOP5

1975748:	/cgi-bin/jouhou/protherm/pp_entry.pl (個別エントリーの CGI)
181190:	/cgi-bin/jouhou/protherm/pp_mutation.sh (mutation 表示の CGI)
57532:	/jouhou/Protherm/protherm.html (ProTherm のホーム)
55529:	/cgi-bin/jouhou/protherm/protherm.pl（データ検索の CGI）
20163:	/jouhou/Protherm/protherm_search.html (検索画面)

４）ProNIT の月別アクセス統計

月:　　リクエスト総数: ページリクエスト数
Apr 2009:　46287:　　1758:
May 2009:　51344:　　2227:
Jun 2009:　69372:　　2456:
Jul 2009:　61098:　　2164:
Aug 2009:　74397:　　2121:
Sep 2009:　85434:　　2654:
Oct 2009:　90104:　　3040:
Nov 2009:　85664:　　2655:
Dec 2009:　63512:　　2410:
Jan 2010:　70534:　　2571:
Feb 2010:　60444:　　2624:
Mar 2010:　94635:　　3456:

５）ProNIT のアクセスドメイン:TOP10

1195619: 64.30%: .com (Commercial)
156604: 11.31%: .jp (Japan)
162279: 10.52%: [unresolved numerical addresses]
128726:　8.67%: .net (Networks)
15577:　1.90%: .edu (USA Higher Education)
17270:　1.01%: .tw (Taiwan)
2364:　0.31%: [unknown domain]
3232:　0.29%: .in (India)
2302:　0.23%: .de (Germany)
1101:　0.21%: .org (Non Profit Making Organisations)

６）ProNIT のリクエストページ：TOP5

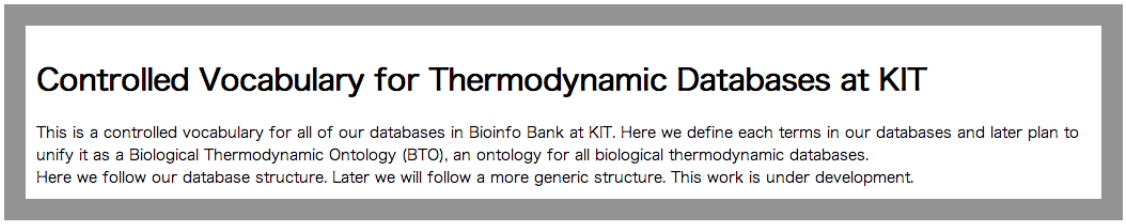
1513842:　/cgi-bin/jouhou/pronit/new/bind_entry.pl　（個別エントリーの CGI）
52013:　/jouhou/pronit/pronit.html　(ProNIT ホーム)
11662:　/jouhou/pronit/pronit_search.html　(検索画面)
15332:　/cgi-bin/jouhou/pronit/new/pronit_search.pl　(データ検索 CGI)
4737:　/cgi-bin/jouhou/pronit/new/pronit_stat.pl　(Statistics の CGI)

別紙参考資料２
蛋白質･蛋白質相互作用データベースのプロトタイプ

entrynumber
protein information:
protein1 name
protein1 synonyms
protein1 source
protein1 sequence
protein1 biologicalunit
protein1 uniprot
protein1 pdb
protein1 mutation
protein1 asa
protein1 secstr
protein1 prothermnumber:
protein2 name
protein2 synonyms
protein2 source
protein2 sequence
protein2 biologicalunit
protein2 uniprot
protein2 pdb
protein2 mutation
protein2 asa
protein2 secstr
protein2 prothermnumber
Complex information:
pdb complex
ligand
conformation
Experimental Condition:
Temperature
pH
BufferName
BufferConcentration
Additives
protein1 Concentration
protein2 Concentration
Ion Name
Ion Concentration
Method
Binding Data:
Kd
Kd Mutant
Ka
Ka Mutant
dG
dG Mutant
dH
dH Mutant
dCp
dCp Mutant
Stoichiometry
Literature:
Reference
Author
Keywords
Remarks
RelatedEntries

別紙参考資料３

熱力学データの Controlled Vocabulary



Controlled Vocabulary for Thermodynamic Databases

Expand All　Collapse All

- ProNIT
  - Protein Information
  - Nucleic Acid Information
  - Complex Information
  - Experimental Details
  - Binding Data
  - Literature
  - General
- ProTherm
  - Sequence and Structural Information
  - Experimental Details
  - Thermodynamic Data
  - Literature
  - General

別紙参考資料４

（１）XML フォーマットによる ProNIT データの一部

```
<pronit xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:noNamespaceSchemaLocation="file:pronit.xsd">
 <entry>
  <entryNumber>1</entryNumber>
  <proteinDetails>
   <name>Myb proto-oncogene protein</name>
   <synonyms>c-Myb protein; Transforming protein myb</synonyms>
   <source>Mus musculus (Mouse)</source>
   <fragment>89-193</fragment>
   <sequence>MARRPRHSIYSSDEDDEDIEMCDHDYDGLLPK</sequence>
   <biologicalUnit>1</biologicalUnit>
   <dbReference type="PIR">TVMSMB</dbReference>
   <dbReference type="SWISSPROT">MYB_MOUSE (P06876)</dbReference>
   <dbReference type="PDBFREE">1MBE, 1MBG, 1MBJ</dbReference>
   <dbReference type="PROTHERM">786 787 788 789</dbReference>
   <mutation>wild</mutation>
  </proteinDetails>
  <nucleicAcidDetails>
   <name>MBS-I (22-mer)</name>
   <source>Synthetic</source>
   <type>DDS</type>
   <sequenceTopWild>caccctaactgacacacattct</sequenceTopWild>
   <sequenceBottomWild>agaatgtgtgtcagttagggtg</sequenceBottomWild>
   <mutation>wild</mutation>
   <dbReference type="GENBANK">M33654</dbReference>
  </nucleicAcidDetails>
  <complexDetails>
   <dbReference type="PDBCOMPLEX">1MSE</dbReference>
   <dbReference type="PRONUC">86</dbReference>
   <proteinConformation>R2 and R3</proteinConformation>
   <nucAcidConformation>The base pairs </nucAcidConformation>
  </complexDetails>
  <exptDetails>
   <temperature>20.2 C</temperature>
   <pH>7.5</pH>
   <bufferName>Potassium phosphate</bufferName>
   <bufferConc>100 mM</bufferConc>
   <ionNameOne>Potassium chloride (KCl)</ionNameOne>
   <ionConcOne>20 mM</ionConcOne>
   <method>Isothermal titration calorimetry (ITC)</method>
  </exptDetails>
  <bindingData>
   <Kd_Wild>5.00e-08 M</Kd_Wild>
   <Ka_Wild>2.00e+07 1/M</Ka_Wild>
   <dG_Wild>-1.21e+01 kcal/mol</dG_Wild>
   <dH_Wild>-1.25e+01 kcal/mol</dH_Wild>
   <dCp_Wild>-6.20e-01 kcal/mol/K</dCp_Wild>
   <stoichiometry>1.01</stoichiometry>
  </bindingData>
  <citation>
   <reference>J Mol Biol. 1998; 276(3):571-590 PMID: 9551098</reference>
   <author>Oda M, Furukawa K, Ogata K, Sarai A, Nakamura H</author>
   <keywords>c-Myb; DNA-binding; ITC</keywords>
  </citation>
  <miscellaneous>
   <remarks>R2R3* (C 130 I), a stable mutant</remarks>
   <relatedEntries>2, 3, 4, 5, 6, 7, 8, 9, 10, 11</relatedEntries>
  </miscellaneous>
 </entry>
 <copyright>copyright</copyright>
</pronit>
```

（２）フラットフォーマットから XML フォーマットに変換する手順

１．フラットファイルからデータ構造の情報を取得

２．この情報から XML スキーマを定義

３．XML スキーマの情報をプログラムにインプリメント

４．フラットファイルのすべての項目の読み込みと修正

５．フラットファイルから HTML 情報部分を削除

６．フラットファイルから項目とデータ列を分離

７．各列を読み込み CVS ファイルに書き出し

８．CVS ファイルを読み込みブランク部位は削除

９．CVS ファイルからデータを抽出し XML スキーマに従って XML ファイルを生成

１０．XML 構文をチェックし XML ファイルをバリデート

1. Get the data structure from the flat file
2. Define the XML schema based on this structure
3. Incorporate the XML schema structure information into the program
4. Read the flat file and check all the fields. If there is any error, correct it.
5. Remove the HTML content from the flat file
6. The flat file contains two columns. Split the columns
7. Read the columns and convert the column data into a CSV file
8. Read the CSV file and remove the fields having null data
9. Extract the data from the CSV file and create the XML file as per the XML Schema
10. Verify the XML syntax and validate the XML file.

別紙参考資料５

九工大補完課題プロジェクト専用の Web ページ

整備実績一覧

（１）保有データ情報

（１－１）データの種類

①生物種	複数
②試料・ライブラリー等の種類、数	①蛋白質およびその変異体の構造安定性に関する熱力学データ約 24,700 件 ②蛋白質と核酸の相互作用の定量的な熱力学実験データ約 10,400 件
③測定方法	熱測定、分光測定など
④データの内容	書誌情報、数値情報、画像情報。内容の詳細は成果報告書別紙参考資料参照。
⑤その他、特記事項	なし。

（１－２）データソース

①現在のデータ量	
②データ区分	■ 自前　■ 第三者　■ 文献データ　■ 計算結果等の二次データ □その他（下欄に詳細を記述）
③将来の増加の見込み	蛋白質の安定性および蛋白質・核酸相互作用の熱力学データともに、年間 2,000 件程度の新規データが発生。
④権利関係	所有者（データベース作成者及び九州工業大学） 公開（■可　□否　□その他［　　］）
⑤その他、特記事項	**①ProTherm: http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html** **②ProNIT: http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html**

（１－３）データの管理状況

①更新頻度等の管理状況、体制	月に１回程度更新。データベースサーバーは無休でアクセス可能。
②その他、特記事項	なし。

（１－４）データベース関係

①DB 管理者数	１名
②キュレータ･アノテータ数	２名
③データ構造	リレーショナル
④DB 管理ソフト	SYBASE
⑤サーバの OS	Linux
⑥サーバ規模	ワークステーション
⑦DB へのアクセス数	**ProTherm:** 年間約 5 万件、**ProNIT:** 年間約 4 万件
⑧独立 IP 数	約１万個
⑨その他、特記事項	DB の検索メニューの画面コピーは別紙参考資料添付。 オントロジーは今後整備。

（２）データ（又はＤＢ）の連結、統合化整備（※試験的、限定的公開済みのものも含む。）

通番	データ（又はＤＢ）の名称	公開／未公開	概要（データの種類（生物種）・数量（kB 等）、本プロジェクトで実施した特徴点、進捗状況、今後の計画・課題などを簡潔にわかりやすく記述）
1	ProTherm	公開	1998 年より公開。利用状況の詳細は成果報告書別紙参考資料参照。構造データベース PDBj とクロスレファレンスにより連結。今後もデータベースの更新と構造データベースとの統合を継続。人材の確保が課題。
2	ProNIT	公開	2001 年より公開。利用状況の詳細は成果報告書別紙参考資料参照。構造データベース PDBj とクロスレファレンスにより連結。今後もデータベースの更新と構造データベースとの統合を継続。人材の確保が課題。

（３）ＤＢ基盤システム、ツール等開発成果物の整備（※試験的、限定的公開済みのものも含む。）

通番	ＤＢ基盤システム、ツール等の名称	公開／未公開	概要（主な機能・特徴点、進捗状況、今後の計画などを簡潔にわかりやすく記述）
1	XML 化の変換プログラム	公開 （XML データは公開）	ProNIT のデータを XML 形式への変換し公開している。

（４）その他の成果物（（２）、（３）に該当しないもの）

通番	名称	公開／未公開	概要