平成１９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」

１９年度　研究成果報告書

平成２０年３月

国立大学法人九州工業大学　情報工学部　　教授　皿井　明倫

成果報告書

委託業務の題目

「生体分子の熱力学データと構造データの統合」

実施機関

住　所　福岡県北九州市戸畑区仙水町１－１
機関名　国立大学法人 九州工業大学
担当者　皿井明倫

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１９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

これまでに収集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22,000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する。これにより、熱力学データベースの任意のデータから対応する構造データにリンクするとともに、構造データベースの任意の構造から対応するすべての熱力学データにリンクすることができるようにする。

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

これまでに収集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7,800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する。これにより、熱力学データベースの任意のデータから対応する構造データにリンクするとともに、構造データベースの任意の構造から対応するすべての熱力学データにリンクすることができるようにする。

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については、まだ実験データが当機関から生成されていないので、平成１９年度はまず入れ物となるデータベースのプロトタイプの設計を行い試験的に既存のデータを組み込み本格運用の準備を進める。

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

熱力学データについては、これまでにオントロジーやデータフォーマットなどの整備があまり行われていない。そこでまず、オントロジーについて調査を行い、オントロジー構築の準備を行う。同様に、熱力学データについて試験的なXMLフォーマットを作成する。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の作成に着手する。これらの開発に当たっては、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

我々が構築している蛋白質と変異体の熱力学データベースと PDBj が構築している構造データベースを統合するため、これまでに収集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 22,000 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。このクロスレファレンスでは､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からなる。これにより、熱力学データベースの任意のデータから対応する構造データにリンクするとともに、構造データベースの任意の構造から対応するすべての熱力学データにリンクすることができる。PDBj からはすでに構造データから熱力学データベースへのリンクを作成した。このクロスレファレンスの情報は、データベースの更新（通常は月に一度）に合わせて更新するようにした。なお、クロスレファレンスのテーブルは以下の URL からダウンロードすることができる。蛋白質と変異体の熱力学データベースの検索画面の内容については、別紙参考資料を参照。

クロスレファレンスの URL：

http://dna01.bse.kyutech.ac.jp/jouhou/Protherm/Protherm_CrossReference.html

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

我々が構築している蛋白質・核酸相互作用の熱力学データベースと PDBj が構築している構造データベースを統合するため、これまでに収集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 8,500 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。このクロスレファレンスでは、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からなる。

これにより、熱力学データベースの任意のデータから対応する構造データにリンクするとともに、構造データベースの任意の構造から対応するすべての熱力学データにリンクすることができる。現在、PDBj において構造データから熱力学データベースへのリンクを作成中である。このクロスレファレンスの情報は、データベースの更新（通常は月に一度）に合わせて更新するようにした。なお、クロスレファレンスのテーブルは以下の URL からダウンロードすることができる。蛋白質・核酸相互作用の熱力学データベースの検索画面の内容については、別紙参考資料を参照。
クロスレファレンスの URL：

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

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については、現在試行的な実験が進行中で本データはまだ生成されていない。この実験では蛋白質と蛋白質の相互作用の強度などの定量的なデータが網羅的に生成される予定である。そこで、今年度はまず受け皿となるデータベースのプロトタイプの設計を行った（プロトタイプはまだ設計段階なので作成完了次第公表の予定)。このデータベースも蛋白質・核酸相互作用の熱力学データベースと本質的には同じスキーマによる関係データベースを用いており、格納するデータの種類や型などの検討を行った。

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

異種のデータを統合するにあたっては、異なるスキーマやフォーマットのデータを処理する必要がある。したがって、これらの異種データの交換フォーマット、用語、セマンティックスなどの標準化を行う必要がある。蛋白質・核酸相互作用の熱力学データベースについて、プロトタイプとなる試験的な XML フォーマットを作成した。また、これまでのフラットフォーマットから XML に変換を行うためのプログラムを試験的に作成した。一方、これらのフォーマット生成に必要となるオントロジーについては調査を行った。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。しかし、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、すでに存在する他のドメインのオントロジーで熱力学用語を含む例などを調べた。オントロジーの整備においては、個人の努力だけでは困難で、同じ分野での研究者の協力が必須である。本年度は、メリーランド大学で蛋白質・リガンド相互作用の熱力学データベースを作成している Mark Gilson 教授と意見交換を行い、熱力学データのオントロジーの整備を協力して行うことになった。また、PDBj は構造データのオントロジーの整備を行っており、今後情報交換を行いながら協力していく。

データ収集やデータベースの統合化においては、テキストマイニング技術の開発が必要となる。今年度は、熱力学データの掲載された文献を自動収集するための方法や、文献からデータを自動抽出する方法を開発するための準備を行った。文献の自動収集では、最適な検索のキーワードの設定を試みている。データの自動抽出では、統合 DB センターが開発している文献のマーキングツールを試用し、マーキングの候補キーワードリストの準備などを行った。

また、統合 DB センターがすすめているデータベースの横断検索に対応するため、熱力学データのインデックス作成の準備を行った。これまでに、熱力学データベースの更新に合わせてホームページから生データをダウンロードできるようにした。なお、熱力学データ全体のダウンロードについては、知的財産所有権の関係で、許可を得てから行うようになっている*) ので、今回のダウンロードサイトは一般ユーザーへの公開でなく統合 DB センターのみへの公開となっている。

*) 詳細は以下のサイトを参照：
http://gibk26.bse.kyutech.ac.jp/jouhou/Protherm/protherm_copyright.html
http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit_copyright.html

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	19	洋誌	Thermodynamic Database for Proteins: Features and Applications	M. Michael Gromiha and Akinori Sarai	Methods in Molecular Biology				2008	印刷中

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	19	国内	**生体分子間相互作用の熱力学データベースと解析**	皿井明倫	生物物理学会	2007 年 12 月 21 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
	19			

２.４ 活動（運営委員会等の活動等）

中核機関、協力機関との会合の履歴

件名：補完課題の連携についての打合せ
日時：10 月 19 日（金）12：30～2：00
場所：ライフサイエンス統合データベースセンター　センター長室
参加者：九工大：皿井、統合 DB センター：高木（TV 会議)、西川、永井

件名：「蛋白構造に関連した統合」のための打合せ
日時：11 月 7 日（水）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
参加者：理研：豊田、国島、九工大：皿井、蛋白研：中村（TV 会議)、統合 DB センター：高木、西川、文科省：田中

件名：「生体分子の熱力学データと構造データの統合」のための打合せ
日時：2008 年 2 月 8 日 PM1:00～1:40
場所：九州工業大学・大阪大学蛋白質研究所間でのテレビ会議
参加者：九工大：皿井、蛋白研：中村、統合 DB センター：大野

件名：テキストマイニングに関する打合せ
日時：2008 年 3 月 14 日 PM3:30～4:30
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：西川、八塚、山口、大野

２．５ 実施体制

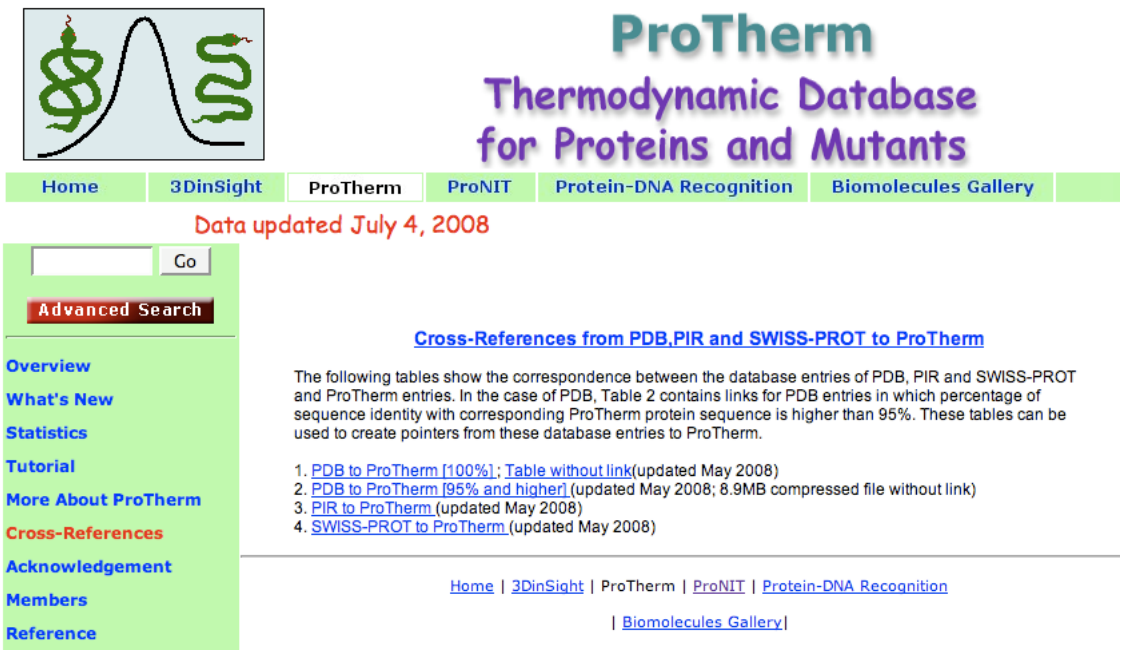
研 究 項 目	担当機関等	研究担当者
1. 蛋白質と変異体の熱力学データベースの構築と統合	九州工業大学情報工学部	◎○皿井 明倫
2. 蛋白質･核酸相互作用の熱力学データベースの構築と統合	九州工業大学情報工学部	○Shaji Kumar
3. 蛋白質･蛋白質相互作用データの生成と統合	九州工業大学情報工学部	○末田 慎二
4. XML データフォーマットやオントロジーなどの統合化技術の開発	九州工業大学情報工学部	○藤井 聡

注1. ◎:課題代表者、○:サブテーマ代表者

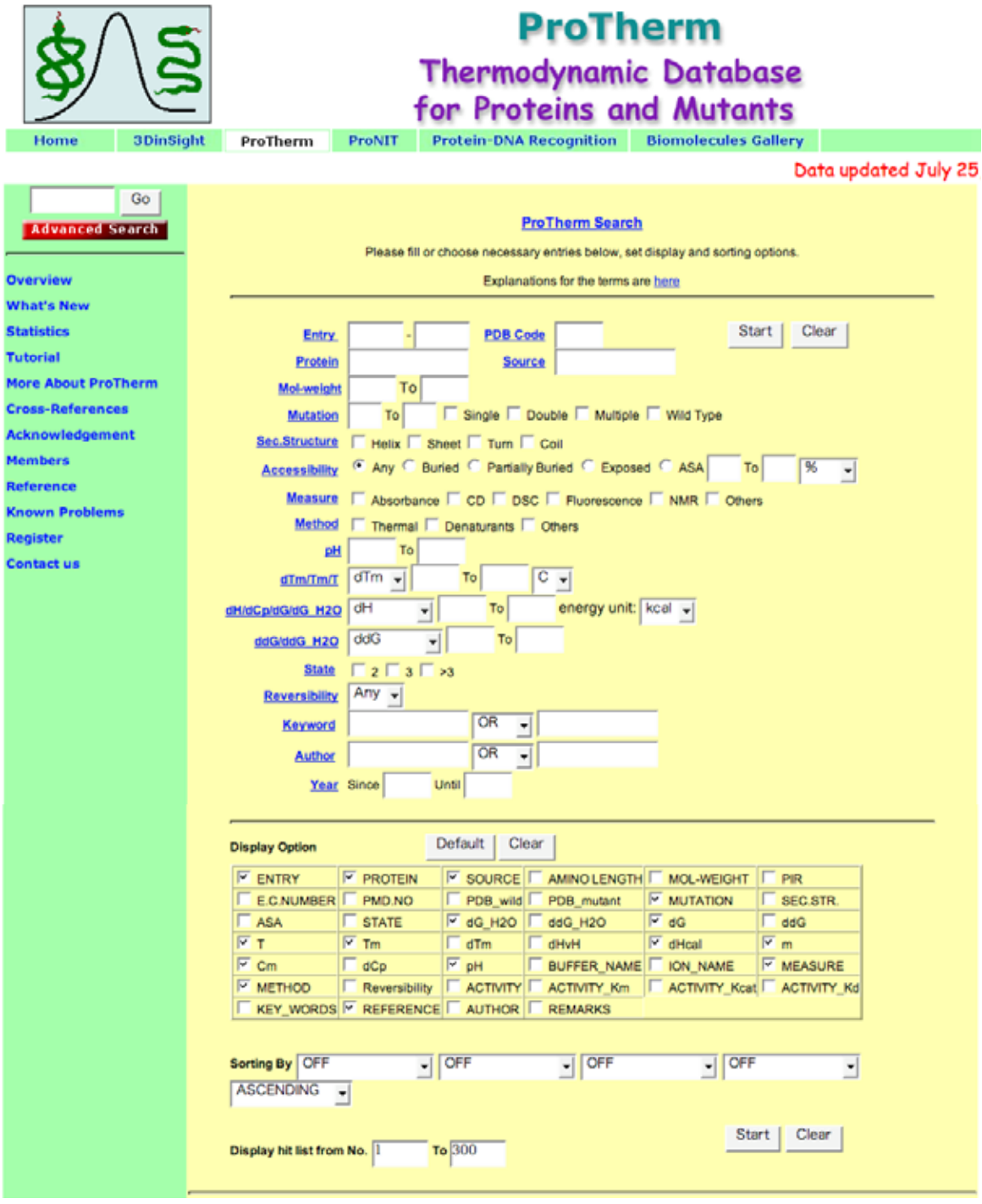
注2. 本業務に携わっている方は、全て記入。

別紙参考資料

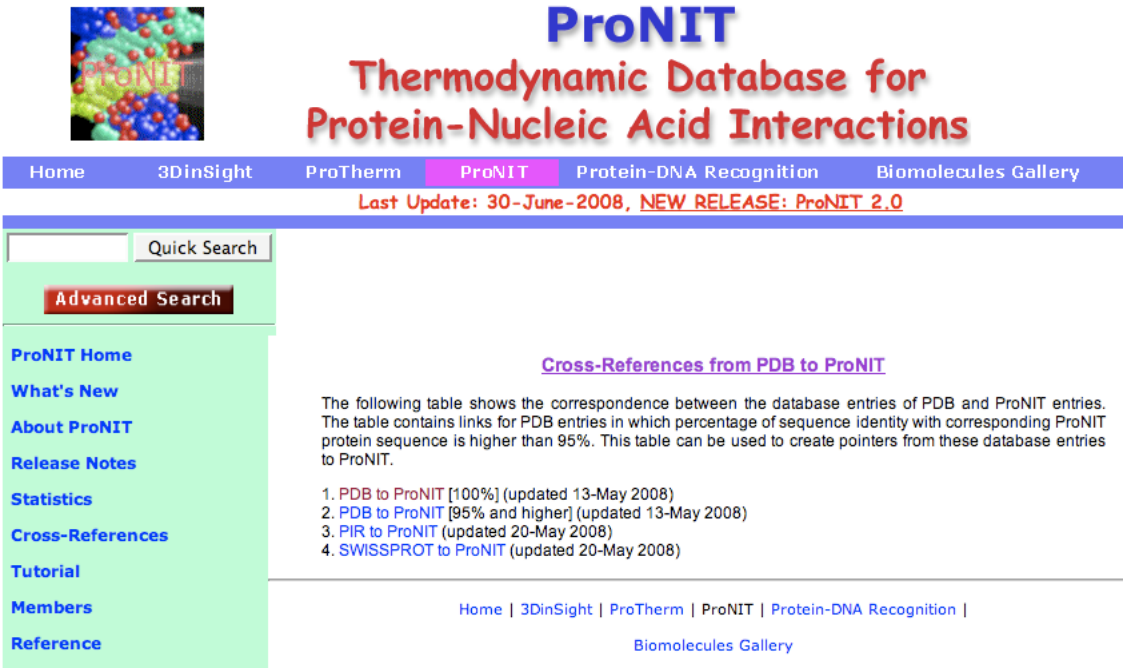
（１） 蛋白質熱力学データベース ProTherm のクロスレファレンステーブル画面



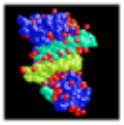
（２） 参考のため蛋白質熱力学データベース ProTherm の検索画面の例を示す。



（３）蛋白質・核酸相互作用熱力学データベース ProNIT のクロスレファレンステーブルの画面



（４）参考のため蛋白質･核酸相互作用熱力学データベース ProNIT の検索画面の例を示す。



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

（５）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。