テクニカルコラム

植物プロモーターデータベース ppdb の紹介

山本　義治、日恵野綾香（岐阜大学応用生物科学部）

本コラムでは岐阜大学山本研究室で運営している植物プロモーターデータベース（ppdb : http://ppdb.agr.gifu-u.ac.jp）について紹介させて頂きます。

DB 自前開発のいきさつ

2003 年にプロモーターを構成する機能配列を一括抽出する方法を思いついたのがそもそもの発端です。

2001 年当時著者のひとり（山本）は理研ゲノム科学センターに在籍していました。篠崎研のリソースを用いて初めてマイクロアレイ解析を行い、そのデータを用いて当時「出来るはずだ」と思われていたアレイデータからの未知の転写制御配列の予測（*ab initio* 予測）にトライしていました。既存のアプローチとしてコンセンサス配列の抽出による予測法が既に確立されていましたが、例えばその頃の Jeff Dangl の発表（Gordon 会議？）では既存の方法はいずれも成績が悪く信頼の置ける予測法ではないという評価であり、マイクロアレイデータをもとにした新規の転写制御配列の予測は「時期尚早」ということのようでした。

コンセンサス配列の抽出による予測法はタンパク質のモチーフ検出法が転用されたものであり、個人的な印象（山本）では「スジが悪い」と感じられました。この方法ではある応答を示す大半のプロモーターがひとつのコンセンサス配列を持つ、という状況が仮定されていますが、生理学的に考えるとそのような均一な応答系は想像しづらく、また、転写制御配列ではなくてもコンセンサス配列となりうる例が多々ある、というようなことを考えていました。

そこで、自前で別のアプローチによる予測法を作りましたが、これだと「うまくいきそう」な感触は得られたものの予測精度がいまひとつでした。状況改善のためには別の独立した解析により「転写制御配列であることの保証」があると役に立つことがわかってきました。

その頃読んだ論文のひとつに、ヒトのいくつかの転写制御配列の翻訳開始点を基準にした出現場所を調べたところ、プロモーター上流から下流へ向けて分布パターンの偏りがあった（Elkon et al., 2003）という報告がありました（図 1）。この論文に刺激を受け、逆に、「出現に局在性を持つ配列」をすべて抽出すればプロモーターの機能配列が網羅的に得られるのではないか、と考えました。そのアイデアに沿って新規のプロモーター構成因子抽出法を開発し LDSS 法（Local Distribution of Short Sequences）と名付けました。プログラム類やデータが一通り揃い論文をまとめている際に、このアプローチには前例がある（FitzGerald et al., 2004）ことがわかりがっかりしてしまいましたが、そうはいってもそこはそれ、いわゆる「植物では初めて」ということで気を取り直して論文にしました（Yamamoto et al., 2007b）。これは山本にとっては初の純バイオインフォマティクスの論文ですが、Faculty of 1000 Biology では EXCEPTIONAL という評価を得ることができました（自慢です、すいません）。

抽出された配列群はこれまでのところプロモーター構成因子としては最も網羅的なものであると自負しています。これらをグループ化すると、TATA box、転写開始点のコンセンサス配列である Inr、Kozak 配列、高等植物固有の新規コアプロモーター因子であると考えられる Y Patch と GA Element、CA Element、そして既知の転写制御配列を含む約 300 の配列群、に分けられました。これらのプロモーター構成因子をゲノム配列にマップしていけばプロモーター構造を認識することが可能になります。

抽出された LDSS 陽性配列の利用法としては、これらの配列をゲノムへマッピングし個々のプロモーターの構造を表示する、というのが最も有用です。そこで、ヒトプロモーターデータベースを作成している方やデータベース作成の経験のある方などに上記抽出配列を用いたプロモーターアノテーションとそのデータベース化をそれとなく打診してみたのですが、返って来たのは「自分でやれば」という大変冷たい御返事ばかりでした。皆さん自分のデータは自分で DB 化する、というスタンスなようです。そこで仕方なく、なく泣く自前で DB 化を行うことにしました。幸い当時参加していたゲノム特定の支援班のサポートを得ることができたおかけで無事に DB が完成し、2007 年 6 月に公開へとなんとか漕ぎ着けることができました（Yamamoto and Obokata, 2008）。最初のバージョンはシロイヌナズナとイネに対応しています。その後 2 度アップデートを行い、ヒメツリガネゴケデータと新機能（プロモーターの種間比較ができる）を追加しました。

ということで私にとっては馴染みのない分野での活動であったため論文化、DB 作成の際にはいろいろと戸惑った記憶があります。発端となったマイクロアレイデータからの転写制御配列の予測法については 2009 年に岐阜大学に移動してから研究を再開し最近方法論として確立することができました（Yamamoto et al., 2011）。

ppdb の特徴

ppdb から得られるデータと操作性に関する特徴を以下にリストアップしました。

- ✓ シロイヌナズナ・イネ・ヒメツリガネゴケに対応
- ✓ 詳細な転写開始点情報が得られる *
- ✓ コアプロモーター（TATA、Inr、Y Patch、GA Element、CA Element）の構造・タイプがわかる *
- ✓ 転写制御配列（REG）の情報が得られる *
- ✓ REG 配列から遺伝子グループへ逆検索が可能
- ✓ 種間オルソロググループの並列解析が可能
- ✓ TAIR GBrowse とデータリンクしている

* オリジナルデータに基づく

次項で具体的な利用例を紹介しますが、遺伝子 ID からプロモーター情報（転写開始点、コアプロモーター構造・タイプ、REG）を呼び出すだけでなく、特定の REG に注目して該当する遺伝子リストを参照することもできます。また、種間オルソロググループから複数の遺伝子を選択することによる並列解析も可能です。ちなみにシロイヌナズナの情報については TAIR の GBrowse からデータリンクされています（ppdb を表示する設定にする必要あり）。GBrowse を見ていて、表示されている TSS やプロモーターエレメントのアイコンをクリックすると ppdb へ誘導され、該当遺伝子が表示されます。

データベースのもとになっている情報は以下の 4 点です。

① ゲノム配列と遺伝子モデル
（TAIR、RAP-DB、COSMOSS から）

② 転写開始点情報
（シロイヌナズナ：理研完全長 cDNA（RAFL）の 5′ 末端情報及び CT-MPSS により自前で同定した 160K の TSS タグ情報（Yamamoto et al., 2009）、イネ：KOME 完全長 cDNA の 5′ 末端情報、ヒメツリガネゴケ：基生研長谷部らによる 1M の 5′ SAGE 情報）

③ LDSS 法により得られたプロモーター構成因子
転写制御系配列である REG（Regulatory Element Group）に加えて TATA Box、Inr、Y Patch、GA Element、CA Element のコアプロモーター因子群がゲノムごとに抽出されている（Yamamoto et al., 2007a ; Yamamoto et al., 2007b）。

④ PLACE 情報
肥後ら（Higo et al., 1999）によりまとめられた転写制御配列情報

② と ③ はオリジナルデータであり、④ は ③ の注釈として用いられています。ppdb を参照しなければ得られない情報としてはやはりオリジナルな ② と ③ ということになります。

利用法

遺伝子 ID からプロモーター情報を呼び出すほかに、特定の REG に注目して該当する遺伝子のリストを参照することができます（図 2）。

1. 植物種を選び調べたい遺伝子の ID を「Keyword search」へ入力、GO のカテゴリーから遺伝子を選択する。
2. プロモーター構造を表示して TSS、コアプロモーター構造・タイプ、REG の位置や種類などの情報が得られる。
3. 特定の REG（転写制御配列候補）をクリックすると該当遺伝子リストを逆検索できる。

REG はトップページにグループ化してあるので数字をクリックして呼び出したリストから特定の REG へたどることもできます。

さらに種間オルソロググループの任意の遺伝子について並列解析ができます（図 3）。

1. 植物種を選び調べたい遺伝子の ID を「Homologue Gene Search」へ入力、GO のカテゴリーから遺伝子を選択する。

2.　種間オルソロググループを呼び出し、チェックを入れた遺伝子のプロモーター構造を並列に表示できる。

利用上の注意

注意点その 1

デフォルトではプロモーター構成因子（LDSS 陽性配列）の表示が制限されています。LDSS 陽性配列には期待される出現位置のエリアがあり、初期設定ではその領域内のものだけが表示されます。例えば TATA box なら－45 から－18 までの位置にある場合にのみ表示されます（主要転写開始点（**Peak TSS**）からの距離）。

表示を制限すると見やすくてよいのですが、問題点としては、① ひとつの遺伝子についてプロモーターが複数ある場合最も主要なプロモーター以外のものの構造は表示されない、② 転写開始点情報がない遺伝子（相当ある）についてはプロモーター構造が全く表示されない、ということがあります。これらの問題を解消するには表示制限をキャンセルする必要があります。具体的には、配列が表示されているウィンドウ（**Focused view**）の下にある“**All**”ボタンを押すとこの出現位置による表示制限は解除されます（図 4）。このボタンは“**Reliable**”<=>“**All**”のトグルになっています（デフォルトでは“**All**”が表示されています（＝押すと **All** になる）。

注意点その 2

REG（転写制御配列）は遺伝子発現情報を一切使わずに出現位置の情報のみから LDSS 陽性配列として抽出されています。REG と PLACE との関連づけを行ってはいますが生物学的な知見の表示は現状では不十分です。

従って頻繁に次のような状況が生じてしまいます、すなわち、① ppdb を使って調べたいプロモーターに含まれている REG の場所と配列がわかった、② でも、それが生物学的にどのような意味があるのかわからない。

上記の問題解消へ向け、現在マイクロアレイデータからの転写制御配列予測とそれらの機能検証の作業を進めています。前者は REG の予測とは別のアプローチによるものであり、予測配列を発現応答と関連づけることが可能です。最近植物ホルモン応答についての予測結果をまとめたので（Yamamoto et al., 2011）、この情報と REG との関連付けを行い、ppdb に表示する作業を行っています。これにより約 300 あるシロイヌナズナの REG のうち 50 程度について発現応答（植物ホルモン応答：AUX、BR、CK、ABA、ET、JA、SA、H_2O_2、乾燥、DREB1Aox の応答）についての情報を提供することができるようになります。

注意点その 3

ppdb にはユーザーの配列をオンデマンドで解析する、という機能はありません。例えば、トマトのプロモーター配列を見て欲しい、という場合には、オフラインで対応します。

次のアップデート（図 5）

現在はイルミナシークエンサーを用いて得られた新たな転写開始点情報（シロイヌナズナ、34 M タグ分）、マイクロアレイデータをもとにした REG へのストレス・ホルモン応答の Annotation 情報、そしポプラゲノムデータの追加作業を行っています。

Elkon, R., Linhart, C., Sharan, R., Shamir, R., and Shiloh, Y.（2003）. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res *13*, 773–780.

FitzGerald, P.C., Shlyakhtenko, A., Mir, A.A., and Vinson, C.（2004）. Clustering of DNA sequences in human promoters. Genome Res *14*, 1562–1574.

Higo, K., Ugawa, Y., Iwamoto, M., and Korenaga, T.（1999）. Plant cis-acting regulatory DNA elements（PLACE）database : 1999. Nucleic Acids Res *27*, 297–300.

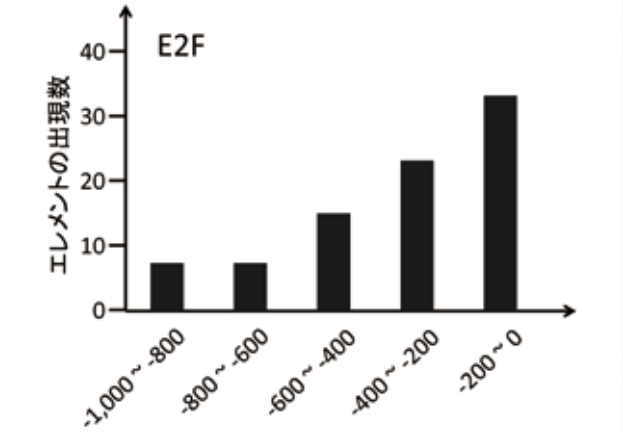
Yamamoto, Y.Y., Ichida, H., Abe, T., Suzuki, Y., Sugano, S., and Obokata, J.（2007a）. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. Nucleic Acids Res *35*, 6219–6226.

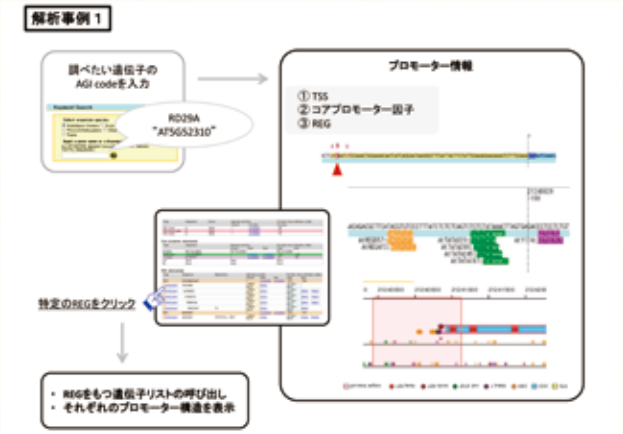
Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T.（2007b）. Identification of plant promoter constituents by analysis of local distribution of short sequences. BMC Genomics *8*, 67.

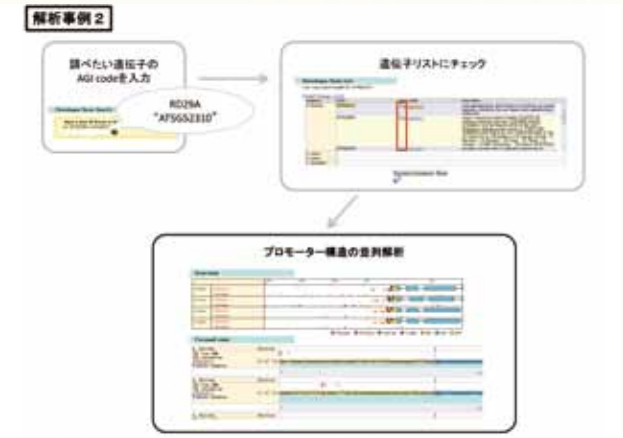
Yamamoto, Y.Y., and Obokata, J.（2008）. ppdb, a plant promoter database. Nucleic Acids Res *36*, D977–981.

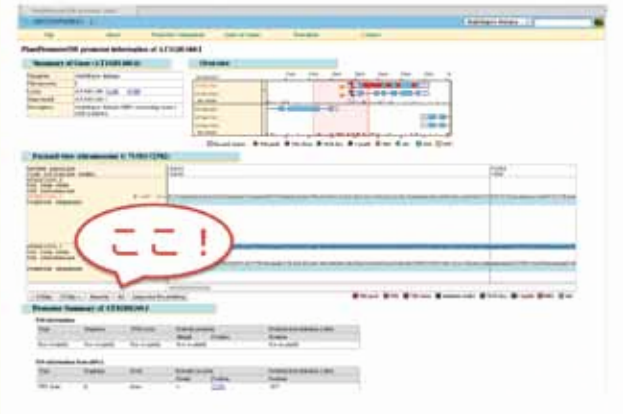
Yamamoto, Y.Y., Yoshioka, Y., Hyakumachi, M., Maruyama, K., Yamaguchi-Shinozaki, K., Tokizawa, M., and Koyama, H.（2011）. Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. BMC Plant Biol *11*, 39.

Yamamoto, Y.Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J.（2009）. Heterogeneity of Arabidopsis core promoters revealed by high density TSS analysis. Plant J *60*, 350–362.









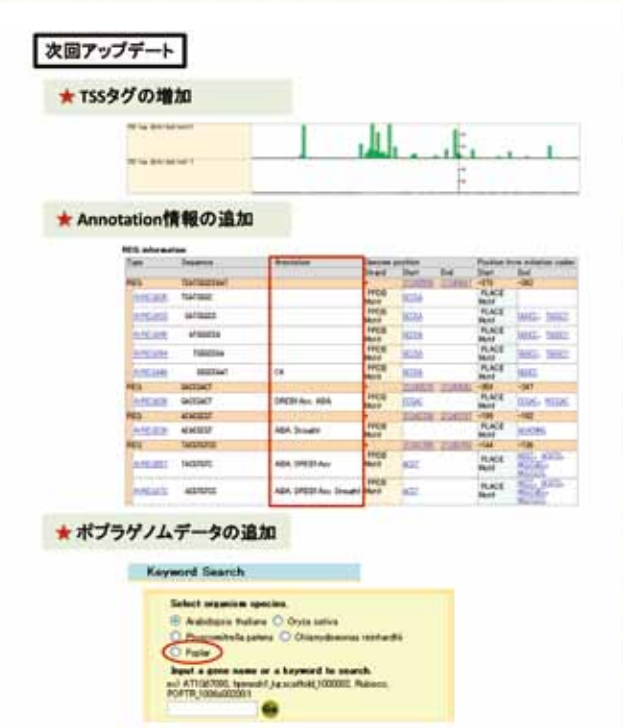


図 1	図 2
図 3	図 4
図 5	

図 1　Elkon ら（2003）の報告による転写因子 E2F の認識配列の分布
グラフの横軸は翻訳開始点からの距離。NRF-1、Sp1、NF-Y、CREB、ATF についても同様の分布を示した。

図 2　解析事例 1
調べたい遺伝子の ID を「Keyword search」へ入力。TSS、コアプロモーター因子、REG の位置や種類などの情報を得ることができる。また、特定の REG をクリックすると該当遺伝子リストを呼び出せる。

図 3　解析事例 2
調べたい遺伝子の ID を「Homologue Gene Search」へ入力。種間オルソロググループから任意遺伝子のプロモーター構造を並列解析することができる。

図 4　プロモーターエレメントの表示制限解除
All ボタンを押すと制限解除される。

図 5　次回アップデート
新たな TSS タグ情報、REG のストレス・ホルモン応答を示す Annotation 情報、ポプラゲノムデータの追加など。