Thermus-Deinococcus 属ゲノムアノテーションデータベースによる新規 DNA 修復遺伝子の検索

久保田千尋　(指導教員：由良敬)

1　はじめに

近年、多くの生物のゲノム塩基配列の決定により、大量かつ多様なデータが得られ、蓄積された大量のゲノム情報をもとに、生物種間の遺伝子の配列や構成を比較するゲノム解析が行われている。異なる生物種間のゲノムを比較することで、生物の多様性が生じた原因となる遺伝子の違いを明らかにすることが可能であろう。しかし、比較解析に用いるゲノムのデータは膨大であるため、有用な情報を効果的に引き出すためには、コンピュータを用いた大規模な情報解析技術の活用が不可欠である。つまり、解析に用いるのに必要なデータを取り出したアノテーションデータベースを開発することで、生物種間のゲノムデータの比較がしやすくすることが必要である。本研究では、２種の細菌、*Deinococcus radiodurans* [1]と *Thermus thermophilus* [2]およびその近縁種のゲノムアノテーションデータベースを開発した。

2　前提知識

2.1　DNA 修復機構

DNA は生物の遺伝情報を保持する生体高分子である。DNA は二重らせん構造になっている。DNA の情報は、RNA に転写された後翻訳され、タンパク質となり生体内で機能を果たす。DNA は放射線照射で損傷する。DNA の損傷は、正常な代謝活動に伴う過程や紫外線の照射などでも発生しているが、細胞には DNA を修復する機構が備わっているため、正常な機能を維持できている。DNA の修復には数多くのタンパク質が関与している。生物の DNA 修復効率は生物種の違いによって大きく異なっている。

2.2　放射線耐性細菌

Deinococcus radiodurans は、放射線照射に対して非常に強い細菌である。この高い放射線耐性は、放射線による DNA 損傷(二本鎖切断)に対して、優れた修復能力を持つことに起因する。しかし、*D. radiodurans* の高効率な DNA 修復機構の詳細はわかっていない。塩基配列の解析の結果、ゲノム自体に特別な物性はなく、機能が不明のタンパク質が多く存在することがわかった。そこで、*D. radiodurans* の強力な DNA 修復機構は、機能未知のタンパク質に担われていると考えられている。ゲノムから、*D. radiodurans* の DNA 修復に関与する新規遺伝子とタンパク質を明らかにできれば、放射線耐性のメカニズムがわかるであろう。

そこで本研究では、DNA 修復に関与する新規遺伝子の発見を試みる。そのために、新規遺伝子を発見する方法を開発し、その手法を *Thermus-Deinococcus* 属のゲノムに適用する。

3　手法

3.1　リレーショナルデータベースを構築したゲノム生物種

膨大な情報量をもつゲノムの中から、重要な機能に関係のある遺伝子部分を探すには、ゲノムが類似している近縁種同士のゲノムを比較することが有力な手段となる。*D. radiodurans* の近縁種として、高度高熱菌の *T. thermophilus* が存在する。*T. thermophilus* と *D. radiodurans* は進化的に近縁な種であるのでゲノムが類似しているが、*T. thermophilus* には放射線抵抗性はない。共通祖先由来であるゲノム同士を比較することで、表現型の違いを生み出すメカニズムがわかると考え、まず、ゲノムの比較を行うツールとして、両者を含む近縁種のゲノムに関連するアノテーションデータベースを作成した。生物種、染色体数、遺伝子数、塩基数は図１に示す。

生物種名	染色体数	遺伝子数	塩基数
Deinococcus radiodurans	4	3,181	3,284,156 nt
Deinococcus geothermails	3	3,062	3,247,018 nt
Deinococcus desertii	4	3,451	3,855,329 nt
Thermus thermophilus HB8	3	2,238	2,116,056 nt
Thermus thermophilus HB27	2	2,210	2,097,482 nt

図１： データベースを構築した生物種

3.2　リレーショナルデータベースの作成方法

２種の細菌の塩基配列データは、GenBank　(NCBI; National Center for Biotechnology Information)が提供している公共のデータベースから取得した。GenBank から取得したデータから、CDS(protein cording region :タンパク質の翻訳領域)の情報を取り出し、加工し、アノテーションデータベースを作成した。CDS の遺伝子には各々固有の ID をつけ、以下の７つの情報に関するテーブルを作った。

3.3　テーブルの名称と説明 (　)内はテーブル名

①　遺伝子の塩基配列の情報　(dna)：ID と塩基配列の情報。

②　アミノ酸配列の情報 (amino_acid)：ID と DNA の塩基配列から翻訳されるアミノ酸配列の情報。GenBank から取得したファイルのアミノ酸配列の情報に誤りがあったので、正しいアミノ酸配列に翻訳し直した。

③　遺伝子の開始点、終止点、鎖 (gene)：ID、遺伝子がコードされているゲノム塩基配列上の位置、鎖の情報。

④　プロモータの情報　(promoter)：ID、プロモータの塩基配列、プロモータ塩基配列から連続する 5 塩基を取り出した断片(fragment)、プロモータの位置の情報。

プロモータは、DNA から RNA の合成の開始に関与する特定領域の短い塩基配列で、遺伝子の発現に重要な領域である。共通のプロモータに制御される遺伝子は類似の機能を持ち、機能が似ているタンパク質を生成する。

プロモータ配列は、CDS 開始領域の上流 50 塩基とした。

⑤　ゲノムから推定されたタンパク質の機能 (protein)：遺伝子の塩基配列に BLAST(Basic Local Alignment Search Tool)[3]を用いてアミノ酸配列の機能を推定した。

⑥　生物種名 (species)：ID と対応する生物種名の情報。

⑦　タンパク質の機能の種類 (p_function)：タンパク質の名称および ID(s_id)と対応するタンパク質の機能の情報

3.4 検索の目的と流れ

現在機能が未知である遺伝子の中から、DNA 修復に関与すると予想される遺伝子の検索を行った。プロモータの塩基配列が類似している遺伝子にコードされているタンパク質は関連する機能を持っている場合があることから、プロモータの塩基配列に着目した。検索の手順は以下に示す。

(1) DNA 修復に関与する既知の遺伝子検索

遺伝子の塩基配列から予測されるタンパク質の中で、DNA 修復に関与する既知の遺伝子を検索した。

(2) 類似するプロモータを持つ遺伝子検索

(1)の検索で得た遺伝子のプロモータ配列と類似のプロモータ配列を持つ遺伝子を検索した。

(3) 機能未知の遺伝子検索

(2)の検索で得た遺伝子のうち、機能が未知のタンパク質をコードする遺伝子を検索した。

以上の検索によって、機能が未知の遺伝子の中で、DNA 修復に関与する遺伝子が推測される。

4 実行結果

4.1 DNA 修復に関与する既知の遺伝子検索

protein テーブルと p_function テーブルに対して、以下の SQL 文を用いて検索を実行した。

・検索文

select distinct id from protein,p_function where p_function.function like '%DNA repair%' and p_function.s_id=protein.s_id and expect <=0.0001;

protein のテーブル (一部)

id	s_id	name	length	expect
NC_001263_1257	RAD16_YEAST	DNA repair protein RAD16	790	6e-23
NC_005835_1155	RECN_ECOLI	DNA repair protein recN	553	1e-13

p_function のテーブル (一部)

s_id	function
RAD16_YEAST	DNA repair
RECN_ECOLI	DNA repair

expect は機能推定の信頼度を意味する数値であり、今回は 10^{-4} に設定した。以上の操作により、316 個の DNA 修復に関与する遺伝子の ID が 53.26 秒で検索できた。

4.2 類似するプロモータを持つ遺伝子検索

類似するプロモータは、同一の 5 塩基断片(=fragment)の有無により判別した。4.1 の検索で特定した DNA 修復に関与する既知の遺伝子において、そのプロモータが持つ fragment を検索した。検索の結果、1008 種類の fragment が検索された。

1008 種類の fragment うち、DNA 修復に関与する遺伝子のプロモータに多く存在する fragment を特定した。特定の方法は、各 fragment に対し、式(1)の値が 1.3 以上のものとした。この結果、DNA 修復に関与する遺伝子に多く存在する fragment を 5 つ同定することができた。同定された fragment とそれぞれの式(1)による値を図 2 に示す。図 2 に示した fragment を持つ遺伝子の ID をプロモータのテーブルから検索した結果、773 個の遺伝子の ID が検索された。

promoter のテーブル (一部)

id	sequence	f 1	f 2	strand	fragment	f_region
NC_008025_1205	ggacggcacct･･	1292998	1293047	c	ggacg	1292998
NC_008025_1205	ggacggcacct･･	1292998	1293047	c	gacgg	1292999

$$\log_2 \frac{\text{DNA 修復に関与する遺伝子のプロモータにおける fragment の頻度}}{\text{全遺伝子のプロモータにおける fragment の頻度}} \quad \textbf{(1)}$$

fragment	値
tgtta	1.92…
gttat	1.73…
cgtaa	1.68…
acata	1.44…
tatgt	1.39…

図 2 : DNA 修復の遺伝子に多く存在する fragment と式(1)による値

4.3 機能未知の遺伝子検索

protein のテーブルには、BLAST の検索結果により、機能が推定された遺伝子の情報が登録されている。したがって、機能が未知のタンパク質の遺伝子の ID は、protein のテーブルに登録されていない ID である。4.2 の検索で得た 773 個の遺伝子の ID のうち、protein のテーブルに登録されていない ID を検索した。その結果、219 個の遺伝子の ID が特定された。この遺伝子のうち、図 2 の fragment を多数持っている遺伝子は、DNA 修復に関与する遺伝子である可能性が高いと考えた。219 個の ID の中で、図 2 の fragment を 3 つ以上持っている遺伝子の ID を検索した。検索の結果、図 3 に示すように 4 個の遺伝子の ID が得られた。

ID	生物種	fragment 数
NC_001263_0217	*Deinococcus radiodurans*	6
NC_012526_2289	*Deinococcus desertii*	4
NC_012526_0305	*Deinococcus desertii*	3
NC_001263_0074	*Deinococcus radiodurans*	3

図 3 : DNA 修復機能を持つと推測される遺伝子と fragment の数

5 まとめと今後の課題

本研究の結果、*Thermus-Deinococcus* 属のゲノムアノテーションデータベースを開発し、DNA 修復に関与すると推測される新規遺伝子を検索することができた。今後は、データベースを活用して、*D. radiodurans* と *T. thermophilus* のゲノムを比較し、表現型の違いを生じるメカニズムを推定したい。

参考文献

[1] *White,O.*,et al. (1999) Genome sequence of the radio resistant bacterium *Deinococcus radiodurans* R1, *Science*, **286** (5444), 1571-1577.

[2] Henne,A., et al. (2004) The genome sequence of the extreme thermophile *Thermus thermophilus*, *Nature Biotechnol.*, **22** (5), 547-553.

[3] Altschu,l S.F., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Research, **25** (17), 3389-3402.