

DGENE ファイルのホモロジー検索用パッケージ GETSIM に関する FAQ

1. Q: GETSIM パッケージでホモロジー検索する場合のシステム制限値を教えてください。

A: 配列質問式の最短配列長 : 5

1 回のコマンド行に入力できる最大文字数 (コマンド、スペースも含めて) : 256

/SQP フィールドでオンライン検索できる最長配列長 : 750

/SQN フィールドでオンライン検索できる最長配列長 : 500

バッチ検索で実行できる最長配列長 : 1,500

SDI 検索で実行できる最長配列長 : 1,500

候補回答数の上限 : 10,000

オンラインホモロジー検索の検索継続時間 : 2 時間 / 1 回まで

2. Q: 500 残基 (核酸) / 750 残基 (タンパク質・ペプチド) より長い配列質問式を実行する場合はどうすればよいですか?

A: 配列長が 1,500 までなら、バッチ検索で実行できます。1,500 を超える場合は、主要な部分配列を選んで 1,500 以下に限定してバッチ検索してください。

3. Q: QUERY コマンドで配列質問式を作成する際に、"EXCEEDS MAXIMUM FIELD LENGTH, WILL BE SEARCHED AS 'XXXXX..." というメッセージが表示されることがあります。これはどういう意味ですか? また、この場合はどうすればよいのですか?

A: QUERY コマンドで配列質問式を作成する場合、配列コードの後に検索フィールドを付加しておかないと、入力した配列コードは基本索引中のテキストデータと仮定されます。ところが、基本索引のテキストデータ長にはシステム制限があります。このため、制限値よりも長い配列コードを入力すると、「この検索フィールド (基本索引) で実行できる単語の長さを超えてしまいました。この質問式は実際には 'XXXXX...' として検索されるでしょう。」というメッセージが表示されます。この場合、例えば /SQP のような検索フィールドを配列コードの後に付加しておけば、このようなメッセージは表示されません。また、このようなメッセージが表示された配列質問式でも、ホモロジー検索に利用することができますし、正しく検索されます。

```
=> que mysflcilplllllascllsysfleqskcrqleelfpppsclgkgtikerfctyydikke
EXCEEDS MAXIMUM FIELD LENGTH, WILL BE SEARCHED AS 'MYSFLCILPLLLLLASCLLSYSFL
EQSKCRQLEELFPPPSCLGKGTIKERFCTYY'
L1  QUE MYSFLCILPLLLLLASCLLSYSFLEQSKCRQLEELFPPPSCLGKGTIKERFCTYYDIKKE
```

```
=> que mysflcilplllllascllsysfleqskcrqleelfpppsclgkgtikerfctyydikke/sqp
L2  QUE MYSFLCILPLLLLLASCLLSYSFLEQSKCRQLEELFPPPSCLGKGTIKERFCTYYDIKKE/SQP
```

4. Q: ホモロジー検索の際に "Incomplete Search" というメッセージが表示されることがあります。これはどういう意味ですか? また、これを避ける方法はありますか?

A: このメッセージは、候補回答数が制限値の 10,000 件を超え、検索が中断した場合に表示されます。これを避けるためには、回答候補数が 10,000 件未満になるようより特定の配列質問式に修正してください。"Incomplete Search" となった場合の課金は以下のようになります。
オンライン検索の場合 : 検索実行料は課金されません。

バッチ検索の場合：検索実行料は課金されますが、回答呼び出し料は課金されません。

5. Q: Smith-Waterman スコアとは何ですか？ どうすれば自分で計算できますか？

A: Smith-Waterman スコアは、実際の配列と配列質問式のコードを比較し、重み付けされたマッチ値 (match value) とギャップペナルティー (gap penalty) を考慮して、最も類似性の高い配列を決めるために計算される値です。これは手計算で簡単に計算することはできません。ホモロジー検索すると、回答には各配列の Smith-Waterman スコアと、最も類似性の高い配列のスコア、両者のスコア比較のパーセンテージが含まれており、類似性が比較しやすくなっています。

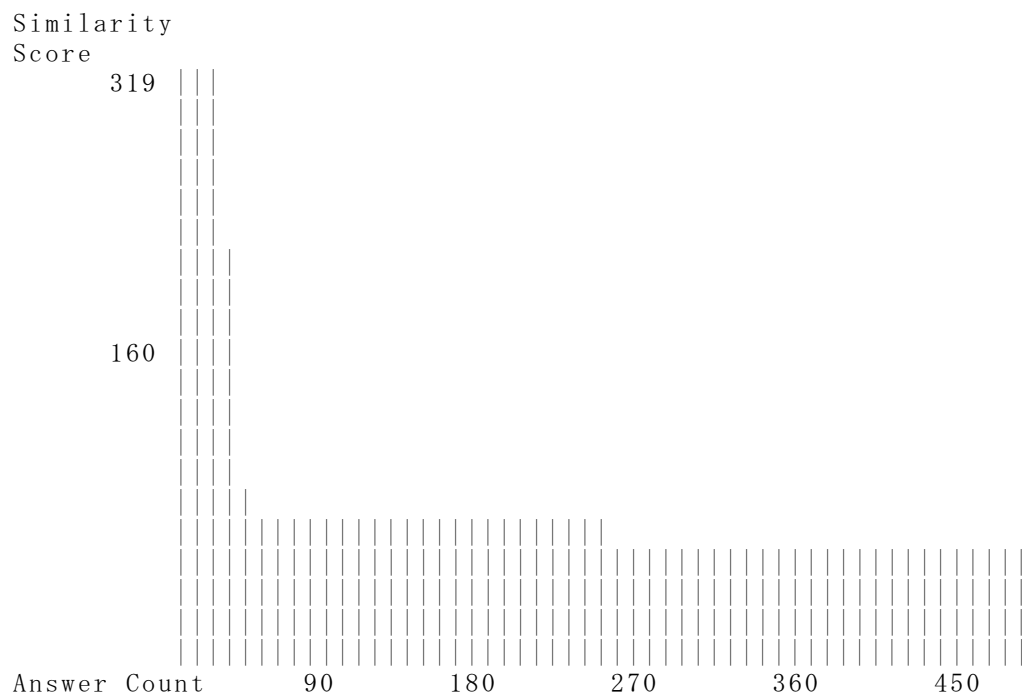
6. Q: 候補回答を検索するためのスコア閾値 (しきいち: threshold) はどのように決定されているのですか？

A: スコア閾値は、適度な回答候補数を調節するために、配列質問式の長さの関数で決定されています。また、検索方法やデータベース規模などに対しても調整されます。例えば、SQP、TSQN、SQN.BOTH などの検索タイプによって別々に決定されますし、データベース規模が大きくなるに従って、変わってゆきます。このように、閾値は変化するものであり、それ自体がホモロジー検索の質を表わしているわけではありません。

7. Q: "QUERY SELF SCORE VALUE" (下の網掛け部分) とは何ですか？ 右の数値は、候補回答中の Smith-Waterman スコアの最高値の場合もありますが、そうでない場合もあるようです。

=> RUN GETSIM L1&L2/SQP
: 途中省略

473 ANSWERS FOUND ABOVE A THRESHOLD OF 50
QUERY SELF SCORE VALUE IS 319



HOW MANY ANSWERS WOULD YOU LIKE TO KEEP ? (ALL) OR ?:

A: QUERY SELF スコアは、配列質問式が回答中の配列に完全に一致した場合に得られる値です。言い換えますと、完全に一致した配列が回答中に存在すれば、その配列の Smith-Waterman スコアが QUERY SELF スコアに一致します。もし、回答中の配列が配列質問式よりも短い場合は、QUERY SELF スコアは Smith-Waterman スコアより大きくなることもあります。

8. Q: 現在のところ、ユーザーはホモロジー検索のスコア閾値 (threshold) を変更することができないように思います。将来、これを可能にする計画はありますか？
A: これについては現在検討しております。

9. Q: 類似した配列の同定パーセント (identity percentage) はどのように計算すればいいですか？

A: ホモロジー検索の回答には、QUERY SELF スコアと、これによって算出された類似性パーセント (similarity percentage) が含まれています。この類似性パーセントは以下のように表示することができます。

```
=> d score  
SCORE 239 15% of query self score 1496
```

類似性パーセントはマトリックスで規定された重み付けを元に計算された値です。これに対して、同定パーセントは純粋に一致しているか否かで計算された値です。この同定パーセントの提示については、現在検討しております。

10. Q: GETSIM パッケージでホモロジー検索を実行すると、回答結果の中に以下の二行が表示されてきます。ここで xxx、nnn、mmmm はそれぞれ数値です。

```
xxx ANSWERS FOUND ABOVE A THERESHOLD OF nnn  
QUERY SELF SCORE VALUE IS mmmm
```

ホモロジー検索がリリースされた当初は、この 2 行目は表示されなかったように思いますが。

A: そのとおりです。2 行目の情報は、ユーザーが類似性を比較しやすいよう Smith-Waterman スコアを標準化するために後で追加したものです。“D SCORE” と表示指定すれば、類似性パーセントも表示することができます。

11. Q: ホモロジー検索した回答レコードを、類似性の高い順に並べてから、特許情報を重複無く表示することはできますか？

A: DGENE ファイルのレコードは、配列単位で構成されています。つまり同一特許由来の配列情報が別レコードに分かれています。このため、回答レコードの特許情報を全件表示すると、同じ特許の情報が重複して表示されてくることがあります。この重複表示を避けるためには、FSORT コマンドで同一特許 (またはファミリー特許) ごとにレコードを並び替えてから “D PFAM” で表示する方法があります。ただし、この FSORT コマンドは、類似性順に並べる “SOR SCORE D” と同時に実行することができません。ですから、ご質問の回答表示は、現在のところできません。

ん。

12. Q: バッチ検索を注文してから、検索待ち (queuing) の段階で取り消した場合、課金はどうなりますか？

A: バッチ検索では、注文した時と回答を呼び出した時に別々に課金されます。ですので、検索待ちの段階で取り消された場合、注文料分は課金されますが、回答呼び出し料は課金されません。

13. Q: UPLOAD コマンドでアップロードするための配列質問式を作成する場合、配列中に空白やキャリッジリターン、改行キーが含まれていても大丈夫ですか？

A: はい。これらの記号は無視して検索してくれます。実際のところ、UPLOAD コマンド用の配列コードは、一行 300 コードまでというシステム制限があり、この範囲内で改行キーを入力しておく必要があります。この制限値を超えてしまっても、アップロード自体は行われますが、検索する際に以下のメッセージが表示されます。

```
WARNING: Your query may have been truncated. A single line in a file
for UPLOAD cannot have more than 300 characters.
DO YOU WISH TO CONTINUE ? (NO): END
```

この場合は、NO (または .、または END) を入力して、検索を取り消してください。

14. Q: GETSIM パッケージでホモロジー検索を実行する前に、アップロードした配列質問式を表示確認する方法はありますか？

A: UPLOAD コマンドでアップロードされた配列質問式は、“D L# LQUE” コマンドで表示確認できます。

15. Q: 配列関連の特許をモニターしたいと考えています。GETSIM パッケージのホモロジー検索を SDI 登録することはできますか？

A: はい可能です。DGENE ファイルでは、GETSIM パッケージのホモロジー検索は “RUN GETSIM 配列コード/検索フィールド ALERT” コマンドで SDI 検索を登録することができます。また、RUN パッケージ (GETSIM、GETSEQ) 以外の質問式も SDI コマンドで登録できます。

16. Q: ホモロジー検索の SDI 検索は、オンラインホモロジー検索と比較して何か相違点がありますか？

A: SDI 検索では検索対象が追加レコードのみであるため、オンライン検索よりシステム制限値が緩和されています。バッチ検索と同様に長い配列質問式 (1,500 まで) を検索できます。また、スコア閾値が下がるため、類似性に対する感受性も高くなり、より多くの回答を得ることができます。

17. Q: 類似性マトリックスにはどんなものが使用されていますか？

A: ホモロジー検索を実行する場合、タンパク質・ペプチドと核酸では、別々のマトリックスが使用されています。

タンパク質・ペプチドをホモロジー検索するためのマトリックス

マトリックス

以下の表は、GETSEQ パッケージで配列検索する際に利用できる一般アミノ酸の 1 文字コードと 3 文字コードをまとめたものです。配列中の特殊アミノ酸は、可能な限り関連する一般アミノ酸で表現されますが、そうでなければ X (または XXX) で表わされます。特殊アミノ酸の具体的な説明は、特徴表に表示されることもあります。その場合は、米国特許商標庁によって提案されている (US Official Gazette 1989 年 5 月 16 日) 特殊アミノ酸の略語と同様な形式で表現されます。

コード表

コードの B と Z は、部分配列検索 (/SQSP と /SQSFP) でのみ使用できます。部分配列ファミリー検索 (/SQSFP) では、B と Z は対応する特定のアミノ酸コード、および一般的な B、Z コードにマッチします。

コード表

コードの Asx と Glx は、部分配列検索 (/SQSP と /SQSFP) でのみ使用できます。部分配列ファミリー検索 (/SQSFP) では、Asx と Glx は対応する特定のアミノ酸コード、および一般的な Asx、Glx コードにマッチします。また、HELP SQQ に例示されている方法で配列質問式に柔軟な条件付けを行うこともできます。

核酸をホモロジー検索するためのマトリックス

マトリックス

文字列の相関関係は IUPAC/IUB の規則に則っています。

核酸配列データのコードは以下の塩基を表わしています。

コード表

化学修飾されている塩基は、親の塩基か N コードで表現され、特徴表に説明が追加されています。その場合は、化学修飾の状態が一般に使われている略語で表現されています。

核酸をアミノ酸配列の質問式でホモロジー検索する場合は、核酸配列をアミノ酸配列に翻訳するために、Universal Genetic Code に基づく翻訳表が使用されています。その検索自体は、タンパク質・ペプチドのホモロジー検索の手順で実行されます。このとき、IUB 記号を採用している DGENE ファイル中のあいまい記号も考慮されています。

コード表

マトリックス

/TSQN の核酸ホモロジー検索で使用するマトリックスは、ストップコドンの情報が追加されたペプチド・タンパク質ホモロジー検索用のマトリックスが使用されています。