

JOIS-Fにおける非文字情報

志村和樹、早瀬修一、相馬 融

日本科学技術情報センター

日本科学技術情報センターが提供しているJOIS-Fには日本化合物辞書データベース、質量スペクトルデータベース、熱物性データベース、DNAデータベースが載っておりオンラインサービスが行われている。また新たに結晶構造、金属材料強度、化学物質法規の3つのデータベースについてもサービス開始に向けてシステム開発とデータチェックが進行中である。ここでは文字情報を扱っている化学物質法規データベースを除く全てのデータベースについて取り上げ、そこで扱われている非文字情報とその取扱について報告する。

NON-CHARACTER INFORMATION IN JOIS-F

Kazuki SHIMURA, Shuichi HAYASE and Toru SOMA

The Japan Information Center of Science and Technology

5-2, Nagatacho 2 chome, Chiyoda-ku, Tokyo 100 Japan

JOIS-F presented by the Japan Information Center of Science and Technology (JICST) gives Chemical Dictionary Database, Thermophysical and Thermochemical Database, and DNA data-base through On-line network. In addition to these databases system development and datacheck on three more databases about Crystal structure, Metallic material strength, and Chemical substance regulation are in progress. All of these databases except for Chemical substance regulation are reported from the view point of non-character database and its handling.

1. はじめに

日本科学技術情報センター(JICST)では、わが国におけるファクトデータベース構築の推進をはかるべく1978年より実験的データバンクシステムの開発に着手し、ファクトデータベースの作成を行ってきた。それらのうち、化合物辞書、熱物性、質量スペクトルの3データベースは、1988年1月よりJICSTファクトデータベースシステム(JOIS-F)の下に一般サービスが開始された。同年10月からはDNAデータベースも加わり、今後は、化学物質法規データベース、結晶構造データベース、金属材料強度データベースが順次追加されていく予定である。(図1)

これらのデータベース間では、物質情報については、化合物辞書データベースに登録され、その際付与された物質識別番号としての化合物辞書番号により、渡り検索を行うことができる。

JOIS-Fとの会話は、公衆電話回線網またはDDX網に接続することによって開始することができる。使用可能な端末種は、公衆回線端末機(パソコン等)の文字端末およびグラフィック端末である。各種パソコン用にソフトメーカーより発売されているグラフィックエミュレータを用いることにより、化合物構造等をグラフィック表示させることが可能となる。

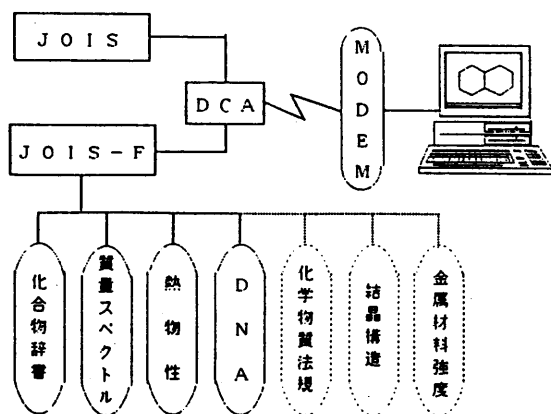


図1 JOIS-Fの構成

2. 化合物辞書検索システム

2. 1 概要

化合物に関する基本的な情報として、体系名、慣用名、分子式、構造式といったようなものが挙げられる。ある対象としている物質の名称は分かっているが構造が分からないとか、逆に構造は判っているのだが名前が知りたい、といった局面がファクト系のデータを扱う場合、共通的な要求として発生する。このような要求に答える為に中心的存在として化合物辞書データベースが存在する。

化合物の表記法として名称、化学式、線型表記(Wiswesser Line Notation等)といった方法が従来より行われてきたが、近年構造情報をトポロジカルな結合表(connection table)の形で表現することが一般的となった。このように化合物をトポロジカルな表現で保持したものが既にファイルとしていくつか提供されているが、この結合表の概念自体はほぼ同様なもので、平面構造を規定するに留まる。だが化合物は本来三次元構造を有しており、立体配置の違いにより化学的性質が大きく異なる場合もある。そこでこのデータベースでは三次元的構造の違いまでも結合表に表現するように拡張を行った。

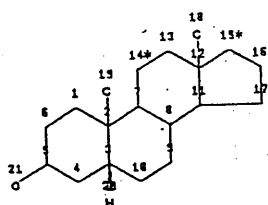
現在約20万件の化合物が蓄積されているが、これらのデータの結合表はすべて名称から自動的に発生される。入力される名称は基本的にIUPACの命名法に準じており、入力された文字列を構文解析し、解析されたフラグメントに対する部分構造情報を再構築することにより化合物全体の結合表を生成する。この時名称に立体記述子(R, S等)があればそれも結合表中に反映されることになる。

2. 2 構造検索

オンラインによるデータベース中の化合物検索の方法として、端末から文字列入力による名称系検索と構造図入力による図形検索に大別される。非文字データの処理として後者が該当するわけだが、図形検索といってもグラフィック図形のパターンマッチではなく、前述の結合表の二次元マトリックスのマッチングを行う。従って端末からは

この結合表を入力しなければならないのだが、その簡便な方法として、端末画面上に検索しようとする構造図を作図する。作図の方法として結合関係を”1-2-3-4-5-6-1”といった数字で指定する手法と、マウス等で端末画面に直接図形を描き、その座標データをホストコンピュータに送信し、ホスト側で結合関係を解析して結合表を作成する2つの手法をサポートしている。

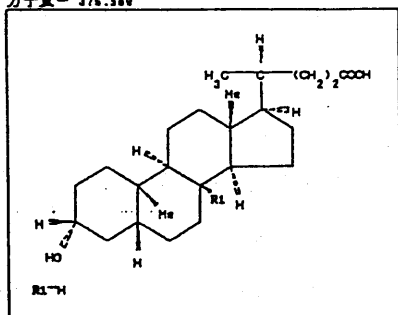
結合表同士のマッチングは非常に処理時間がかかる処理なので二段階の検索に分割して行う。まず作成された質問構造図から構造的特徴をシステムが選出し、全データベースの中からその条件を満たさない化合物を篩い落とす検索を行う。これをスクリーンサーチと呼んでいる。次に残された候補化合物に対して原子対原子の1対1の対応マッチングを行う。検索の為に入力された質問構造図と部分構造検索の回答結果の例を図2に示した。



S: 作図コマンドをどうぞ
U: END

化合物検索システム

000001
SN=J1.532F (C)
ローケータ=BL (バイオロジカルDB)
RN=434-13-8
分子式=C₂₄H₄₀O₃
分子量= 376.589



体系名=3α-ヒドロキシ-5β-コラン-24-酸
慣用名=リトコール酸
慣用名=Lithocholic acid

図2 構造検索例

3. 質量スペクトルデータベース

化合物に関するファクトデータとして各種のスペクトルがあるが、その一つとして質量スペクトルデータをオンライン提供している。このデータベースは質量スペクトルデータベースとして世界的に定評のあるNIST (National Institute of Standards and Technology) のデータと日本質量分析学会の協力により国内の研究者から提供されたオリジナルデータの2つを収録しているが、この両者のデータ形態を統合してデータベース化しているのでファイルの区別を意識せずに検索可能である。

データ内容としては、化合物の一般情報(名称、化学式等)、スペクトルの測定条件等の文字情報と、スペクトルピークの数値情報とがある。ピーク情報の検索は、質量数としてのMzと相対イオン強度RIの対で行われるが、両者とも数値データとして保持されているので範囲指定による検索が可能である。特に質量スペクトルの場合、相対イオン強度は測定条件にかなり依存するので範囲検索が必須の条件となる。このシステムでは、端末からの入力値の±10%の範囲でイオン強度検索を行うように工夫されている。更にこの許容範囲もユーザが変更できるようになっている。

4. 熱物性データベース

4.1 概要

本データベースでは、物質の熱物理学・熱化学的物性データをはじめ、輸送論的性質、光学的・電磁気学的性質、燃焼・爆発に関する性質のような熱力学と関連の深いデータをも含めて蓄積している。

システム開発には、熱物性データベース分科会(故大杉治郎主査)の協力を得、またデータ収集に関しては、(財)生産開発科学研究所および(社)化学工学会の協力を仰いでいる。

対象とする物質・系は元素、無機化合物、有機化合物の純物質及びこれらの物質からなる二・三成分系である。収録可能な物性種は約60物性にのぼるが、このうち現在サービスを行っているもの

```
#0001 TH087000001 JICST COPYRIGHT
C(1)= J34.789B(C) 2-(2,4-DICHLOROPHENO
MW ASSUMED IN EQUATION = 207.05632
STAT= L S
LOG(10)VP = A(1)+A(2)/(T+A(3)) (VAPOR PRESSURE)
IN UNIT OF VP (Torr)
T (C)
A( 1.00)= 7.27816
A( 2.00)= -2.035.515
A( 3.00)= 160.401
RANGE OF T = 211.48: 286.30 C
DOC = 87000001
GRAD= 0 SELECTOR= KAKO
```

図3 式データ出力例

は33物性となっている。

入力データは、前記専門データセンターにおいて科学技術論文等から採択されたものと、権威あるデータブックから利用許諾を得てその一部を収録したものである。

オンライン検索用ファイルは、検索要求の多様性に対応するためにデータ構造に合わせて、物質、物性、書誌の三つのサブファイルから構成されている。

4. 2 数値データの処理

熱物性データの検索では、物質情報からのその物性データの調査以外にも、逆に特定の物性値を持った物質を探す、といったデータ集等の印刷物では多くの労力を必要とする作業を可能にすることが望まれる。このためには、数値情報からの検索、特に浮動小数点に対する検索機能を構築しなくてはならない。

熱物性データベースでは、以下のような数値検索が可能である。

①完全一致

例：TM=273.15

完全に一致するデータを検索するので、この例では標準沸点が273.1500...Kのデータのみが検索される。従って、浮動小数点数に対する完全一致検索は漏れの出る可能性がある。

②指定値以上・以下の検索

例：D>=1000

③指定値より大、未満の検索

例：HM>400.0

④指定値以上かつ以下の検索

例：MW 100 TO 120

⑤近似検索

例：TB?=400

指定した有効数字の最後の桁から1を引いた値より大かつ1を加えた値未満という指定となる。この例では標準沸点が399Kより大かつ401K未満のデータを検索する。

物性値検索の結果、該当するデータが無い時にはシステムが自動的に範囲を拡大して検索を行う機能も持っている。

一方、数値データそのものの保有・表現に関しては、有効桁数と単位の問題がある。測定データの場合、1.0と1.00は値としては同じであっても、その意味するものは異なっている。また、273.15Kと0.00℃は表現形が違うが、内容的には同一のものである。これらの取り扱いは、本データベースでは、次のようになっている。

数値は原典の有効数字情報を保持させ、表示の際も有効桁数を考慮する。したがって、下の桁の余分な0表示は行わない。

会話開始時の検索・表示は、原典での表記法に関わらず、SI単位系に統一している。しかし、他の単位系を使いこなしている利用者のために単位系の自動変換を可能にした。SI系以外にもメートル工学系等の単位系を選択することができる。データの表示も検索もこの選択された単位系で行われる。

また、数値データに対する利用者援護機能として、ヒストグラムの表示が可能である。

文字型データに対してはエキスパンド機能によるデータ通覧ができるが、数値データに対しても指定された幅でのデータの度数分布による概略表

示が可能である。

さらに、一般的なポイント数値だけではなく、範囲データや式データの蓄積・検索もサポートしている。

原典自体に100~110というような幅を持って表記されているデータも、上限値・下限値の両方を保持させ、表示も両方させている。幅を持った場合の検索では、検索指定値の範囲と蓄積物性値の範囲が一部でも重なれば、条件が合致したことになる。

また、登録式データと自由形式の式を集積できる。例えば、蒸気圧におけるAntoinの式を想定した次の形の登録式である。

$$\log_{10}Y=A(1)+A(2)/(T+A(3))$$

この式の係数A(1~3)の値が納められ、表示出力される。さらに、パラメータ（この場合T：温度）の一定の刻み幅で目的変数Y（この場合は蒸気圧）の値を計算し、あらかじめ格納してある。そのため、他の数値データと同様に物性値からの検索を行うことも可能である。式データの出力例を図3に示した。

5. DNAデータベース

1970年代の後半より、世界中の研究者の精力的な研究と塩基配列決定法の進歩により、DNAの塩基配列データが加速度的に蓄積されつつある。この塩基配列自体は“A”，“C”，“G”，“T”の4つのキャラクタの列であるが、場合によっては十数万の文字列で構成されるものもある。

J I C S TのDNAデータベースでは、欧州分

子生物学研究所が収集しているE M B Lと米国のロスアラモス国立研究所が収集するG e n B a n k[®]の二つのデータベースをオンライン提供しているが、この両者の合計は既に5万エントリーを越えている。このような膨大な塩基配列の中から必要とする配列の相同性を調べるためには、大型コンピュータの助けを必要とする。

前述したように塩基配列は単純な文字列であるが、これがアミノ酸の配列をコードしているのであり、コードとしての意味から塩基配列の機能や進化の問題などについて考える際には、ホモロジー検索という手法が要求される。

ホモロジー検索とは、端末から入力される塩基配列とデータベース中の塩基配列との文字列によるストリングサーチではなく、相方の配列中に挿入または飛び越しを許しながら且つ最も相同性の高い部分を見つけ出す、高度な検索手法である。まずデータベース中の塩基配列を横軸に、テスト配列を縦軸に置くドットマトリックスを作成し、その対角線上で一致している部分の多い所を選び出し、そのマッチングポイント（Initial Score）が一定値以上のものについて最適並置の処理に進む。最適並置は、対角線を中心に前後左右一定のウィンドウに領域を拡大し、この中で挿入と欠損を考慮しながらマッチングポイント（Optimized Score）を計算することにより行われる。このシステムでのホモロジー検索結果の例を図4に示す。アルゴリズムとしては、Lipman-Pearsonの方法とGotohの方法をもとに行っている。

図4 ホモロジー検索回答例

```
[ 3] U: RESULT
SUBNO IDENTIFIER      I-SCORE  O-SCORE
G01 HUMGRP5E          112        120

1  TTCAAAGATGTAGGTTCAAAAGGCAAAGGTAAGA
   :::::::::::::::::::: :
410 TTCAAAGATGTAGGTTCAAAAGGCAAAGTTGGTAGA

      LENGTH =      36, MATCHES =      32 ( 88.8 %)
S: 出力を終わりました
```

6. 結晶構造データベースシステム

6.1 概要

我々を取りまく日常製品にはたくさんの結晶性材料が使用されている。各種の全自動家電製品に組み込まれているマイクロプロセッサに使われているシリコンの単結晶、時計には欠かせない水晶、及びドライヤーの温度調節に使用されているセラミクス（多結晶性材料）等、まさに生活に密着している。

また、医薬品の成分の多くは分子性結晶である。医薬品の効能はその成分分子の立体特異性に依存する場合が多く、その分子設計において成分分子を構成する原子の絶対配置の情報は重要な役割を果たす。

これらの材料の特性は、単結晶としての物性あるいは生理活性そのものを利用しているものが多い。従ってこれらの材料の改良、さらには新しい材料、新しい物性の開発には、結晶構造に関する系統的な情報しかも物性的な情報とリンクさせることのできる情報が必要である。結晶構造データベースは、この様な要求に答える道具になることを目指して作られたものである。

6.2 結晶構造データの入力、処理、出力

6.2.1 データの入力

収録データのもととなる原論文の採択は、現時点ではJICST科学技術文献ファイルから行なっている。また採択された原論文からのデータ作成は、日本結晶学会の協力により行なっている。結晶学会を通して記入者には原論文とデータシート（またはMS-DOSフォーマットのフロッピーディスク）が送られ、記入者は決められたフォーマットのデータシート中の記入項目を原論文を読みながら埋めていく。データシートを使う記入者は鉛筆を使って、フロッピーディスクを使う記入者はデータ入力支援システム（データ入力専用のエディター）を使ってデータ項目を記入する。

記入（入力）データ項目として最も重要なものは、結晶構造の3次元グラフィック表示に必要な以下の3つのデータである。

（1）非対称単位中の原子座標

（2）格子パラメータ

（3）空間群

6.2.2 データの加工と貯蔵

記入（入力）済みのデータは、データシートまたはフロッピーディスクの形でJICSTに納品される。納品後のデータ加工の流れを図5に示す。データシートで納品されたデータは前処理をした上で外注パンチに出し、フロッピーディスクとパンチリストの形で納品される。但しここで納品されたフロッピー中のデータ・ファイルはIBMフォーマットの8インチディスク中でのテキスト形式になっている。JICST内でのパソコンを使ったデータ処理は全てMS-DOSの標準テキスト形式で行っているのでフォーマット変換を行う。このようにしてMS-DOSの標準テキスト形式のファイルとして集められたデータはハードディスクに蓄積し次の作業に回す。この段階でデータのバックアップをMTに落とす。

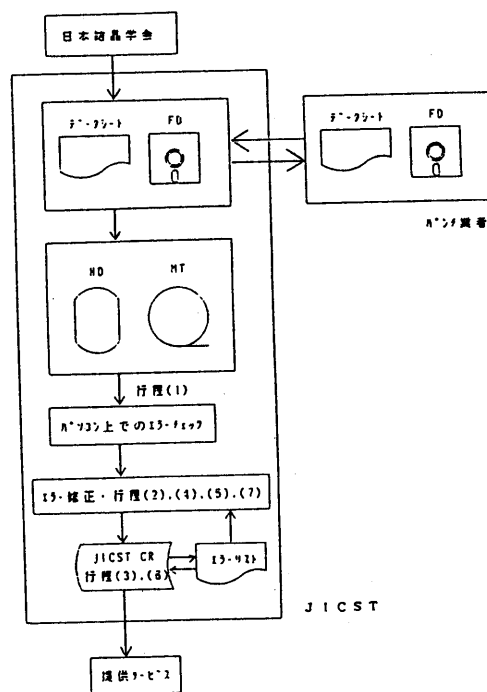


図5：納品後のデータ加工

ここからデータの修正作業に取り掛かる。修正作業は以下のような行程で行う。

(1) パンチミスの発見とその修正(市販のスクリーンエディターを使用)。

(2) パソコン上でのデータチェックプログラムを使ってエラーを見つけ修正する。

(3) ここまでの修正が終わったデータを再びIBM形式に変換して大型計算機のディスクに落とし入力チェックを行う。

(4) (3)の結果、出力されるエラーリストを見ながらデータ修正を行う。

(5) (4)と平行してパソコン上で結晶構造のグラフィック表示用のデータ形式の構造データを、納入されたデータからつくる。この段階でパソコン上では結晶構造のグラフィック表示が可能になるので、これを見ながら次の段階のチェックを行う。

(6) (4)の段階の修正済みデータを再びIBM形式に変換して大型計算機のディスクに落としマスターチェックを行う。

(7) (6)の結果出力されるエラーリストを見ながらデータ修正を行う。

(8) データ項目の記入もれをチェックする。もれがあれば記入する。

(9) マスターファイル作成を行う。

以上のような修正作業の結果として、データの蓄積が行われる。結晶構造そのもののデータは、数値情報として各種記録媒体に蓄積されていくことに注意して欲しい。

6.2.3 結晶構造データの出力

図形形式の回答出力例を図6に示す。原子が存在する位置が点で示され、結合が破線で示されている。注目すべきは、原子の座標データが数値情報から図形情報(視覚情報)に変化するのがこの段階であるということである。図6の出力例は、オンライン処理による出力であるが、ユーザー各自が自前で用意したアプリケーションを利用することで出力形態を変えることができる。と言うのは各ユーザーは原子のグラフィック表示に必要な情報を数値情報としてユーザー端末に出力させることができるので、その出力の中にある情報

を手持ちのアプリケーションを利用して自由に加工できるからである。ただし現時点では、データベース自体に座標データのダウンロード機能はサポートされていない。

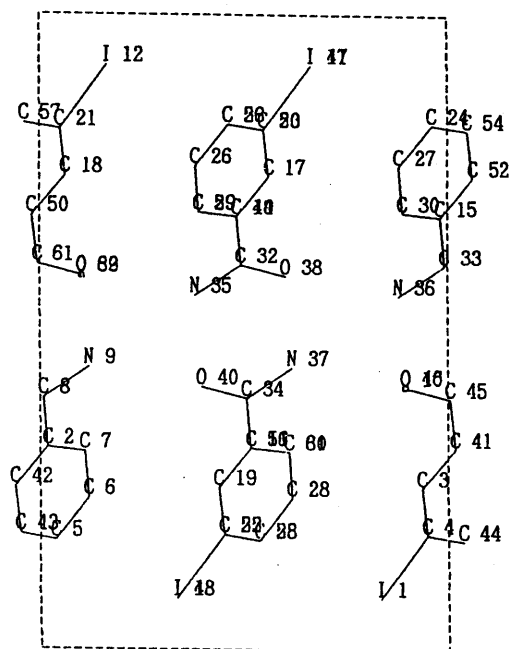


図6 結晶構造オンライン出力

7. 金属材料強度データベース

7.1 概要

金属材料強度データベースは、1985年度より、J I C S Tと科学技術庁金属材料技術研究所(N R I M)との間で共同研究を行い開発を進めてきたものである。材料選択、機械・構造物の設計、安全基準の設定、品質保証、寿命・余寿命推定の利用へ供するために、一般サービスを来年第一四半期中に開始する予定である。

データとしては、主としてN R I Mで行われてきた鉄鋼材料等のクリープ試験・疲労強度試験による測定データである。これらの材料特性データは、本質的に統計的なばらつきを持っている。そのため、通常の数値処理のみならず、統計解析処

理・その視覚化のためのグラフィック処理が必要となる。

7. 2 統計解析

本データベースでは、オンライン上での統計解析機能として次にあげるものをサポートする予定である。

高温引張試験	高温引張特性	直交多項式
	耐力特性	直交多項式
クリーブ破断試験	S-T特性	TTP法
高サイクル疲れ試験	S-N特性	PROBIT法
低サイクル疲れ試験	S-N特性	PROBIT法
	ϵ -Nf特性	最小二乗法
	ϵ -N25特性	最小二乗法
き裂伝ば試験	dA/dN特性	非線形最適化
疲れ応力ひずみ試験	S- ϵ 特性	最小二乗法

利用者が検索の結果作ったデータ集合に対し、これらの解析法を指定することにより、その場でオンライン解析が行われ、回帰パラメータが求められる。

7. 3 データ・解析結果のグラフィック出力

試験データの数値を数字で表しただけでは人間にとっては理解が容易ではなく、出力時の直感的判断、その後の解釈・利用のためにも、グラフに描かせた方が便利である。本データベースでは、文字によるデータ出力のほかにも、グラフ上へのデータプロット、さらには、前述のオンライン解析結果による回帰線を引かせることが出来る。

表1 座標情報の表現

項番	内 容	ビット位置							
		2 ⁷	2 ⁶	2 ⁵	2 ⁴	2 ³	2 ²	2 ¹	2 ⁰
1	Y座標上位部	YH	P	0	1				YH
2	Y座標下位部	YL	P	1	1				YL
3	X座標上位部	XH	P	0	1				XH
4	X座標下位部	XL	P	1	0				XL

しかし、JOIS-Fのオンライン機能は、公衆電話回線による通信を前提としているため、グラフィックデータの出力にはいくつかの問題を生ずる。

まずデータの表現法であるが、図形そのものをビットマップとして持つのが自由度は高いが現在の300、1200bpsといった通信速度では多くの転送時間を必要とする。そこでグラフィック画面を、線画と文字の集合と考え、どこからどこまで線を引くか、どの位置にどの文字を描くか、といったストローク型データとして取り扱っている。

またデータの転送法においては、JOIS-Fにおいても、文献データベースであるJOISや他の商用データベースで以前から使われている、データ長7bit、偶数パリティというプロトコルをコンパチビリティを考えて使用している。これは元来、文字情報、それもASCII文字を電送することのみが考慮されていたのであり、グラフィックデータのようなバイナリデータを送るにはなんらかの工夫が必要となる。

JOIS-Fではエスケープシーケンス等の機能文字を用いた処理によりこれを実現している。基本的には、テクトロニクス社のグラフィック端末に準拠したものとなっている。例えば、座標情報は表1に示すように、10bitのバイナリデータを5bitずつに分けて文字コード領域にマッピングし、2文字のデータとして送ることになる。グラフ出力のイメージを図7に示す。

8. おわりに

わが国におけるファクトデータベースの必要性が認識され、本格的構築が開始されたばかりである。これから更に発展、充実させてさせていく必要があるだろう。

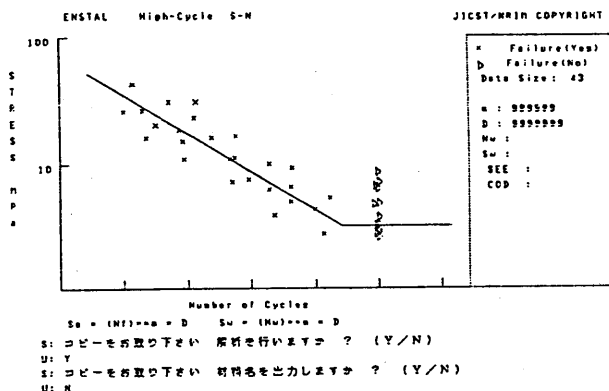


図7 グラフ出力イメージ