Description of Structure of the Folktale : Using a Bioinformatics Multiple Alignment Program

Description of Structure of the Folktale: Using a Bioinformatics Multiple Alignment Program

JUN'ICHI ODA
Institute for the Study of Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies

INTRODUCTION

In a recent paper[1] I produced a rough sketch of an attempt to apply genome informatics (bioinformatics that treats genetic information as a whole) analysis method developed from the field of molecular biology to the study of the folktale. The objective of this attempt was to scientifically reconsider the conventional analogies[2] of the various concepts used for both analysis of various narrative genres and biology at a concrete analysis level. At that time, I raised the following items as the preliminary stages to be performed before serious application could be carried out.

(1) A description of folkloristic text based on Fillmore's case grammar and Jackendoff's semantic structure model.
(2) Setup of a database containing the fundamental materials required for the seme-dictionary (a dictionary with the items that appear on the surface layer of folkloristic text set as headings, that uses articulated semes as reference data for each item).
(3) Application of an alignment program used for genome informatics to the study of the folktale.

An accurate logic model construction exists for item (1), and currently contributions in related areas are under investigation. With regard to the database construction in item (2), data input has been in progress for several years, and official announcement of an experimental system that will allow searching using the web site of the Institute for the Study of Languages and Cultures of Asia and Africa (Tokyo University of Foreign Studies) will be made in the very near future. This paper is mainly concerned with item (3) - investigation of the effectiveness of the alignment programs used in bioinformatics in analysis of folktales by actual application of the software. The folktale sequence structure used as data for the analysis was provisionally based on Propp's[3] 'sequence of functions' model (although the actual model used was a modification of this model by Greimas[4]).

The reason that I presumed to use the Propp model (as of the present a number of problems have been pointed out) is that it is considered that Propp's Functions model produces data suitable for processing.[5] In addition, it was Propp who first made a 'scientific' proposal to use biological analysis methods in folktale research, and more than anything else, I chose his model to pay him homage for this.

ALIGNMENT

With regard to the alignment in bioinformatics, each element of the nucleic acids (4 types) that make up DNA, and each element of the amino acids (20 types) that make up proteins is represented as a character, and hyphens (gaps) are inserted at appropriate locations in the sequences in such a way that the common characters in the sequences are aligned in vertical columns. The following example[6] shows the results of processing on five input protein sequences. An asterisk under a column in the output sequences indicates a maximum similarity value (i.e. all the characters in the column are the same), while a plus sign indicates that the measured similarity exceeds some specified value.

```
..VKQIESKTAF QEALDAAGDK LVVVDFSATW CGPCKMIKPF FHSLSEKYSN VIFLEVDVDD..
..SDKIIHLTDD SFDTDVLKAD GAILVDFWAE WCGPCKMIAP ILDEIADEYQ GKLTVAKLNI..
..ALEEEDNVLV LKKSNFEEAL AAHKYLLVEF YAPWCGHCKA LAPEYAKRAA KLKAEGSEIR..
..DSAVVKLATD SFNEYIQSHD LVLAEFFAPW CGHCKNMAPE YVKAAETLVE KNITLAQIDC..
..METAMSKGVI TITDAEFESE VLKAEQPVLV YFWASWCGPC QLMSPLINLA ANTYSDRLKV..
```

Fig. 1 Sequence input

```
-------VKQIE-SKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFLEV-DVDD
------SDKIIHLTDDSFDTDV-LKADGAILVDFWAEWCGPCKMIAPILDEIADEYQGK--LTVAKLNI
ALEE--EDNVLVLKKSNFEEAL--AAHKYLLVEFYAPWCGHCKALAPEYAKRAAKLKAEG-SEI-R---
------DSAVVKLATDSFNEYI--QSHDLVLAEFFAPWCGHCKNMAPEYVKAAETLVEKN-ITLAQIDC
-METAMSKGVITITDAEFESEV-LKAEQPVLVYFWASWCGPCQLMSPLINLAANTYSDR--LKV-----
                 *             ++ * * *** *+   *     +
```

Fig. 2 Alignment output

When a pattern made up of vertically aligned characters in the aligned sequences is determined to bestow a characteristic, this pattern is called a *motif* for the local similarity. When the shaded section in the above sequence is constructed as a pattern, the following description is obtained.

F[SWYF]A[TEPS]WCG[HP]C[KQ]

The characters in the square brackets are the amino acids permitted for that column. When the similarity relationship between sequences for the above description is analyzed, it provides useful data for the motif. In addition, analysis of various relationships such as specific character insertion and deletion provides a lot of information about the evolution process.

The alignment program used for this paper was CLUSTAL W, Version 1.6 (March 1996)[7], developed by Julie Thompson, Toby Gibson, and Des Higgins. Another program called TREEVIEW[8] was used to display and print the phylogenies from the tree file that outputs the distances between the sequences obtained from the alignment process. CLUSTAL W was run on a Hitachi 9000V Series computer (V715S-Tiny) running HP-UX 10.20, and TREEVIEW was run on a Power Macintosh 6200/75. The input and output formats required for the programs are described in the following chapter.

DATA

The data created for this paper is based on Propp's Function model. Propp proposed the following Function concept based on analysis of Russian fairy-tales (*Narodnye russkie skazki* by A. N. Afanas'ev).

Function is understood as an act of a character, defined from the point of view of its significance for the course of the action.[9] Propp established the following 31 functions.

Table 1 31 Functions by Propp

β : absentation
γ : interdiction
δ : violation
ε : reconnaissance
ξ : delivery
η : trickery
θ : complicity
A : villainy
a : lack
B : mediation, the connective incident
C : beginning counteraction
↑ : departure
D : the first function of the donor
E : the hero's reaction
F : provision or receipt of magical agent
G : spatial transference between two kingdoms, guidance
H : struggle
J : marking
I : victory

K : the initial misfortune or lack is liquidated
↓ : return
Pr : pursuit, chase
Rs : rescue
O : unrecognized arrival
L : unfounded claims
M : the difficult task
N : solution
Q : recognition
Ex : exposure
T : transfiguration
U : punishment
W : wedding

Greimas focused on the Pairs that make up these functions, and based on Propp's catalog, reduced the list to the following 20 functions.

Table 2 20 Functions by Greimas

1: absentation (β)
2: interdiction (γ) vs violation (δ)
3: reconnaissance (ε) vs delivery (ξ)
4: trickery (η) vs complicity (θ)
5: villainy (A) vs lack (a)
6: mediation, the connective incident (B) vs beginning counteraction (C)
7: departure (↑)
8: the first function of the donor (D) vs the hero's reaction (E)
9: provision or receipt of magical agent (F)
10: spatial transference between two kingdoms, guidance (G)
11: struggle (H) vs victory (I)
12: marking (J)
13: the initial misfortune or lack is liquidated (K)
14: return (↓)
15: pursuit, chase (Pr) vs rescue (Rs)
16: unrecognized arrival (O)
17: the difficult task (M) vs solution (N)
18: recognition (Q)
19: exposure (Ex) vs transfiguration (T)
20: punishment (U) vs wedding (W)

The data presented in this paper was produced by using the Greimas catalog. The materials used for data creation were Propp's actual results of analysis of the fairy tales. These results are from analysis performed on 45 fairy tales[10] and from tales within the texts. For analysis, Propp does not treat the tale as a whole, but

rather uses the *moves* that comprise the tale as the unit for analysis. In addition, the following type of simplification was performed for presentation of the analysis results.

1) The first seven functions (β, γ, δ, ε, ζ, η, θ) which are regarded as the preparatory part of the tale are not shown.
2) Trebling (e.g. three tasks, three years of service etc.) is omitted.

For technical reasons relating to data creation, the following simplifications were also used in this paper.

1) Although one function (= *genus*) can be sub-divided into multiple species (e.g. A^1 (kidnapping of a person) and A^9 (expulsion) are species of Function A (villainy)), in this paper only functions (= *genera*) are handled (and not species).
2) The negative result of a function (e.g. the negative result $\bar{F}$ of Function F (receipt of magical agent) indicates that nothing is given to the hero) is also treated as the function.
3) The considerably complicated 'sequence of functions' that is seen as problematic with respect to both the analysis results and Propp's scheme has been omitted from the data.[11)]

After applying these simplifications, the sequences resulting from the Propp analysis are as follows. The numerals at the start show the enumeration of tales according to the collection of Afanas'ev, and the characters after the hyphen are the moves given by the Propp analysis data.

Table 3 Schemes by Propp

93-1	AB↑DEF	106-2	AB↑DE↓
93-2	aBF↑DEF	108	A↑DE↓PrRs
95-1	AB↑DEF↓	112	A↑DE↓PrRs
95-2	aBC↑DEF↓	113	ABC↑DEFG↓PrDERs
97	AB↑DEF↓	114	A↑DEF↓PrRsOQW
98-1	AB↑DEF↓	115-1	A↑EFK
98-2	aBC↑DEF↓	115-2	ABC↑DEFIW
99-1	AB↑DEFT↓	125-1	DEFA↑O
99-2	AB↑DEF↓	125-2	FABC↑HJI↓PrRsLQExUW
100	ADEFMNW	126-1	ABC↑K↓
101-1	AFMNW	126-2	DEFABC↑FHJIOQW
101-2	ABC↑KMNExUW	128-1	aBC↑DEFGJK↓W
102-1	AB↑DEF↓	128-2	AC↑G↓K
102-2	aB↑DEF↓	131	ABC↑HIK↓W
104-1	FaBC↑DEF↓U	132-1	DEABC↑K↓
104-2	aFNMNW	132-2	aCF↑DEFGI↓
105-1	ABCF↓K↓	132-3	AC↑DEF↓GMNOQExUW
105-2	aC↑DEF↓PrRs	133-1	ABC↑DEF
106-1	AB↑K↓	133-2	BC↑DEHIK↓

135 AC↑FHIK↓PrRs
136-1 aBC↑DEKF↓
136-2 ABC↑KF↓
136-3 ACHIK↓PrRs
137-1 ↑DEFACHI↓PrRs
137-2 aBCF↑FMNK↓UW
138-1 aBC↑DEFHIK↓PrRs
138-2 AC↑DEFGK↓W
138-3 aBC↑MNK↓
140 ABC↑DEGFIK↓W
141-1 AF↑DEFGIK↓
141-2 AC↑FGK↓UW
143-1 ACF↑HIK↓
143-2 A↑FHIK↓W
144 aMC↑DEFFGMNTW
145 aBC↑GK↓W
148 ABC↑HIK↓
149 ABC↑HIK↓
150-1 ↑DEaBC↑K↓
150-2 aBC↑HIK↓PrRsUW
151 aBC↑HIK↓
152 A↑HIK↓
153 ABCFHIUW↓
154-1 ↑aDEFG↓
154-2 AC↑FHIUW
155-1 aBC↑DEF↓
155-2 aBC↑DEF↓
155-3 ↑SGWF
155-4 AC↑HIJLQExUW
155-5 ABC↑GKU
156-1 ABC↑DEFGK
156-2 aBC↑DEFGK↓UW
156-3 FAC↑GOMNUW
161-1 ABC↑HIK↓
161-2 aC↑FGHIKF
161-3 aBC↑DEFGIK
161-4 aC↑GHIW
162 ↑DEFABCHIK↓W
163 aFC↑GTW↓
164-1 AC↑K↓
164-2 aFC↑TGIW↓
166-1 DEFA↑GW
166-2 AGTQ
167-1 aDEFKTUW
167-2 ABC↑G
167-3 aBC↑MNW↓
247 A↑GMNW↓Q
244 AFExU
139 ↑DEFGABC↑HIKJW↓
139-2 ADEFGOLMNQTUQW
123-1 DFEAAO
123-2 FABC↑FHIKJ↓L
123-3 aBC↑FK↓ExQUW
123-4 ABC↑FK↓ExQUW

As a preliminary step to processing these schemes using the CLUSTAL W software, they were first converted to the Greimas Functions. In this paper, even if one of the functions in the pairs established by Greimas was present, that function was treated as the relevant pair. This is no more that a provisional operation, because the objective of this paper is to investigate the applicability of the alignment program at a technical level. As a result, using the Greimas model, the Propp schemes are represented as 16 functions.[12)]

The following table shows the Propp functions reduced using the Greimas model to 16 functions, and the corresponding amino acid codes (that are recognized by the CLUSTAL W software).[13)]

Table 4 Conversion table

Functions by Greimas		amino acid codes
A vs a	→	A
B vs C	→	C
↑	→	D
D vs E	→	E
F	→	F
G	→	G
H vs I	→	H
J	→	I
K	→	K
↓	→	L
Pr vs Rs	→	P
O	→	N
M vs N	→	M
Q	→	Q
Ex vs T	→	T
U vs W	→	W

Using this conversion, when the Propp schemes are used as sequence input for the CLUSTAL W program, the following format results.[14] The "P" before the numerals for each sequence (move of the tale) stands for Propp.

Table 5 Sequence input used by CLUATAL W

```
>P093-1
    ACDEF
>P093-2
    ACFDEF
>P093-3
    ADPHW
>P095-1
    ACDEFL
>P095-2
    ACDEFL
>P097
    ACDEFL
>P098-1
    ACDEFL
>P098-2
    ACDEFL
>P099-1
    ACDEFTL
>P138-1
    ACDMKL
>P140
    ACDEGFHKLW
>P141-1
    AFDEFGHKL
>P141-1
    ACDFGKLW
>P143-1
    ACFDHKL
>P143-2
    ADFHKLW
>P144
    AMCDEFFGMTW
>P145
    ACDGKLW
>P148
    ACDHKL
```

>P099-2
ACDEFL
>P100
AEFMW
>P101-1
AFMW
>P101-2
ACDKMTW
>P102-1
ACDEFL
>P102-2
ACDEFL
>P104-1
FACDEFLW
>P104-2
AFMMW
>P105-1
ACFDKL
>P105-2
ACDEFLP
>P106-1
ACDKL
>P106-2
ACDEL
>P108
ADELP
>P112
ADELP
>P113
ACDEFGLPEFP
>P114
ADEFLPNQW
>P115-1
ADEFK
>P115-2
ACDEFHW
>P125-1
EFADN
>P125-2
FACDHIHLPQTW
>P126-1
ACDKL
>P126-2
>P149
ACDHKL
>P150-1
DEACDKL
>P150-2
ACDHKLPW
>P151
ACDHKL
>P152
ADHKL
>P153
ACFHWL
>P154-1
DAEFGL
>P154-2
ACDFHW
>P155-1
ACDEFL
>P155-2
ACDEFL
>P155-3
DGWF
>P155-4
ACDHIQTW
>P155-5
ACDGKW
>P156-1
ACDEFGK
>P156-2
ACDEFGKLW
>P156-3
FACDGNMW
>P161-1
ACDHKL
>P161-2
ACDFGHKF
>P161-3
ACDEFGHK
>P161-4
ACDGHW
>P162
DEFACHKLW
>P163

EFACDFHIHNQW
>P128-1
ACDEFGIKLW
>P128-2
ACDGLK
>P131
ACDHKLW
>P132-1
EACDKL
>P132-2
ACFDEFGHL
>P132-3
ACDEFLGMNQTW
>P133-1
ACDEF
>P133-2
CDEHKL
>P135
ACDFHKLP
>P136-1
ACDEKFL
>P136-2
ACDKFL
>P136-3
ACHKLP
>P137-1
DEFACHLP
>P137-2
ACFLFMKLW
>P138-1
ACDEFHKLP
>P138-2
ACDEFGKLW
AFCDGTWL
>P164-1
ACDKL
>P164-2
AFCDTGHWL
>P166-1
EFADGW
>P166-2
AGTQ
>P167-1
AEFKTW
>P167-2
ACDG
>P167-3
ACDMWL
>P247
ADGMWLQ
>P244
AFTW
>P139
DEFGACDHKIWL
>P139-2
AEFGNMQTWQW
>P123-1
EFEAAN
>P123-2
FACDFHKIL
>P123-3
ACDFKLTQW
>P123-4
ACDFKLTQW

DATA PROCESSING AND DISCUSSION

This chapter deals with the processing procedure for sequence input using CLUSTAL W and TREEVIEW, and the analysis of the alignment output. CLUSTAL W outputs two files after processing the sequence input. These are SOMETHING.ALN (the file that contains the alignment) and SOMETHING.DND (for the dendrogram). In general, alignment transformations are performed repeatedly to approach a sensitive result. There was dispersion in the similarity between sequences in the initial results of the alignment performed on the sequence

input given above 'as a whole', and the sensitivity of the result was quite low. In this paper, the procedure given below was used to group suitable numbers of relatively high-similarity sequences from the alignment output as a whole, and perform alignment on these groups again.

① Alignment as a whole

Alignment was performed on the 93 sequences 'as a whole' until the results converged (three iterations), and the results displayed in groups of relatively high similarity.

```
P155-3    ---------DGWF----------
P166-1    -----EFADGW------------
P125-1    -----EFADN-------------
P137-1    ----DEFACHLP-----------
P105-2    -------ACDEFLP---------
P095-1    -------ACDEFL----------
P095-2    -------ACDEFL----------
P097      -------ACDEFL----------
P132-3    -------ACDEFLGMNQTW----
P136-2    -------ACDKFL----------
P098-2    -------ACDEFL----------
P102-1    -------ACDEFL----------
P104-1    -----FACDEFLW----------
P155-2    -------ACDEFL----------
P155-1    -------ACDEFL----------
P102-2    -------ACDEFL----------
P098-1    -------ACDEFL----------
P115-2    -------ACDEFHW---------
P138-1    -------ACDEFHKLP-------
P133-1    -------ACDEF-----------
P128-1    -------ACDEFGIKLW------
P161-3    -------ACDEFGHK--------
P093-1    -------ACDEF-----------
P138-2    -------ACDEF-GKLW------
P156-1    -------ACDEF-GK--------
P144      -----AMCDEFFGMTW------
P099-1    -------ACDEFTL---------
P136-1    -------ACDEKFL---------
P106-2    -------ACDEL-----------
P141-2    ----ACDFG--KLW---------
P145      ----ACDG--KLW----------
P140      ----ACDEGFHKLW---------
P105-1    --------ACFDKL---------
```

P 1 4 3 − 2 −−−−−−−ADFHKLW−−−−−−−−
P 1 5 3 −−−−−−−ACFHWL−−−−−−−−−
P 1 3 7 − 2 −−−−−ACFLFMKLW−−−−−−−−
P 1 3 2 − 1 −−−−−EACDKL−−−−−−−−−−
P 1 5 0 − 1 −−−−DEACDKL−−−−−−−−−−
P 1 0 6 − 1 −−−−−−ACDKL−−−−−−−−−−
P 1 2 6 − 1 −−−−−−ACDKL−−−−−−−−−−
P 1 6 4 − 1 −−−−−−ACDKL−−−−−−−−−−
P 1 0 1 − 2 −−−−−−ACDKMTW−−−−−−−−
P 1 5 5 − 5 −−−−−−ACDGKW−−−−−−−−−
P 1 6 1 − 4 −−−−−−ACDGHW−−−−−−−−−
P 1 5 6 − 3 −−−−−FACDGNMW−−−−−−−−
P 1 6 7 − 2 −−−−−−ACDG−−−−−−−−−−
P 1 2 8 − 2 −−−−−−ACDGLK−−−−−−−−−
P 1 2 5 − 2 −−−−−FACDHIHLPQTW−−−−−
P 1 5 5 − 4 −−−−−−ACDHIQTW−−−−−−−−
P 1 5 4 − 2 −−−−−−ACDFHW−−−−−−−−−
P 1 2 3 − 2 −−−−−FACDFHKIL−−−−−−−−
P 1 2 6 − 2 −−−−EFACDFHIHNQW−−−−−−
P 1 3 3 − 2 −−−−−−−CDEHKL−−−−−−−−
P 1 3 5 −−−−−−ACDFHKLP−−−−−−−−
P 1 3 6 − 3 −−−−−−−−ACHKLP−−−−−−−−
P 1 6 2 −−−−−DEFACHKLW−−−−−−−−
P 1 4 3 − 1 −−−−−−ACFDHKL−−−−−−−−
P 1 5 2 −−−−−−−−ADHKL−−−−−−−−−
P 1 6 1 − 2 −−−−−−ACDFGHKF−−−−−−−−
P 1 4 8 −−−−−−ACDHKL−−−−−−−−−−
P 1 5 0 − 2 −−−−−−ACDHKLPW−−−−−−−−
P 1 4 9 −−−−−−ACDHKL−−−−−−−−−−
P 1 5 1 −−−−−−ACDHKL−−−−−−−−−−
P 1 6 1 − 1 −−−−−−ACDHKL−−−−−−−−−−
P 1 3 1 −−−−−−ACDHKLW−−−−−−−−−
P 1 3 9 −−DEFGACDHKIWL−−−−−−−−
P 1 2 3 − 3 −−−−−−ACDFKLTQW−−−−−−−
P 1 2 3 − 4 −−−−−−ACDFKLTQW−−−−−−−
P 1 3 8 − 3 −−−−−−ACDMKL−−−−−−−−−
P 1 6 7 − 3 −−−−−−ACDMKL−−−−−−−−−
P 0 9 3 − 2 −−−−−ACFDEF−−−−−−−−−−
P 1 3 2 − 2 −−−−−ACFDEFGHL−−−−−−−−
P 1 4 1 − 1 −−−−−−AFDEFGHKL−−−−−−−
P 1 0 8 −−−−−−−ADELP−−−−−−−−−−
P 1 1 2 −−−−−−−ADELP−−−−−−−−−−
P 1 1 4 −−−−−−−ADEFLPNQW−−−−−−

```
P115-1    --------ADEFK----------
P093-3    --------ADP-HW----------
P247      --------ADG-MWLQ--------
P163      ------AFCDG-TWL---------
P164-2    ------AFCDTGHWL---------
P101-1    ----------AFMW----------
P244      ----------AFTW----------
P100      ----------AEFMW---------
P167-1    --------AEFKTW----------
P104-2    ----------AFMMW---------
P166-2    ----------AGTQ----------
P154-1    -------DAEFGL-----------
P139-2    --------AEFGNMQTWQW----
P123-1    ----------EFEAAN--------
```

Fig. 3 Alignment as a whole

② Division into two groups

The above alignment output was broadly divided into two groups based on similarity. For the first group (Group A), 28 sequences were extracted from the 5th sequence in the above results (P105-2), and for the second group (Group B), 18 sequences were extracted from the 40th sequence (P132-1). When alignment was performed again on these two groups the following results were obtained (Group A converged after two iterations and Group B after one iteration).

```
P128-1    -------ACDEFGIKLW------
P161-3    -------ACDEFGHK--------
P093-1    -------ACDEF-----------
P155-2    -------ACDEFL----------
P105-2    -------ACDEFLP---------
P102-1    -------ACDEFL----------
P097      -------ACDEFL----------
P132-3    -------ACDEFLGMNQTW----
P136-2    -------ACDKFL----------
P102-2    -------ACDEFL----------
P098-1    -------ACDEFL----------
P095-2    -------ACDEFL----------
P104-1    ------FACDEFLW---------
P095-1    -------ACDEFL----------
P155-1    -------ACDEFL----------
P098-2    -------ACDEFL----------
P099-2    -------ACDEFL----------
P115-2    -------ACDEFHW---------
```

```
P 1 3 8 − 1    − − − − − − A C D E F H K L P − − − − − − −
P 1 3 3 − 1    − − − − − − A C D E F − − − − − − − − − − −
P 0 9 9 − 1    − − − − − − A C D E F T L − − − − − − − − −
P 1 1 3        − − − − − − A C D E F G L P E F P − − − − −
P 1 3 6 − 1    − − − − − − A C D E K F L − − − − − − − − −
P 1 0 6 − 2    − − − − − − A C D E L − − − − − − − − − − −
P 1 3 8 − 2    − − − − − − A C D E F − G K L W − − − − − −
P 1 5 6 − 2    − − − − − − A C D E F − G K L W − − − − − −
P 1 5 6 − 1    − − − − − − A C D E F − G K − − − − − − − −
P 1 4 4        − − − − − A M C D E F F G M T W − − − − − −
                          * *
```

Fig. 4 Alignment output of Group A

```
P 1 3 2 − 1    − − − − − E A C D K L − − − − − − − − − − −
P 1 5 0 − 1    − − − − D E A C D K L − − − − − − − − − − −
P 1 0 6 − 1    − − − − − − A C D K L − − − − − − − − − − −
P 1 2 6 − 1    − − − − − − A C D K L − − − − − − − − − − −
P 1 6 4 − 1    − − − − − − A C D K L − − − − − − − − − − −
P 1 0 1 − 2    − − − − − − A C D K M T W − − − − − − − − −
P 1 2 5 − 2    − − − − − F A C D H I H L P Q T W − − − − −
P 1 5 5 − 4    − − − − − − A C D H I Q T W − − − − − − − −
P 1 5 5 − 5    − − − − − − A C D G K W − − − − − − − − − −
P 1 6 1 − 4    − − − − − − A C D G H W − − − − − − − − − −
P 1 5 6 − 3    − − − − − F A C D G N M W − − − − − − − − −
P 1 6 7 − 2    − − − − − − A C D G − − − − − − − − − − − −
P 1 2 8 − 2    − − − − − − A C D G L K − − − − − − − − − −
P 1 5 4 − 2    − − − − − − A C D F H W − − − − − − − − − −
P 1 2 6 − 2    − − − − E F A C D F H I H N Q W − − − − − −
P 1 3 3 − 2    − − − − − − − − C D E H K L − − − − − − − −
P 1 3 5        − − − − − − − A C D F H K L P − − − − − − −
P 1 2 3 − 2    − − − − − F A C D F H K I L − − − − − − − −
                          * *
```

Fig. 5 Alignment output of Group B

It is apparent that the Function C (mediation, the connective incident vs. beginning counteraction) and Function D (departure) chains are common elements in the alignment for both Group A and Group B. However, the three Group A functions (A, E, F) and one Group B function (A) that exhibit high similarity are not marked as complete common elements. This is due to the fact that their various sequence inputs were grouped roughly.

When more than a certain number of sequences are aligned in this way and there is dispersion in the distances between sequences, dividing the alignment output into groups with fewer sequences and performing alignment again is an

effective means of achieving convergence. In addition to this type of two-stage alignment, for specific functions for which micro-level analysis is desired, manually editing the alignment output based on co-occurrence of functions to give a more sensitive result, then running the alignment again may also be considered. In either case, it is clear that grouping the sequences allows more information to be obtained.

A number of analyses for the alignment output given above are shown below.

With the exception of two examples (P104-1 and P144), all of the 28 Group A sequences have the function chain A-C-D at the start. In addition, with the exception of one example (P136-2), Function D (departure) is followed by Function E (the first function of the donor vs. the hero's reactions). Further, with the exception of two examples (P136-1 and P106-2), Function E is followed by Function F (provision or receipt of magical agent). As a result, the chain A-C-D-E-F appears as a motif at the start of 24 of the 28 Group A sequences.

Table 6 Motif of Group A

A (villainy VS lack)
C (mediation, the connective incident VS beginning counteraction)
D (departure)
E (the first function of the donor VS the hero's reaction)
F (provision or receipt of magical agent)

In the case of the 18 Group B examples on the other hand, some have Function E or Function F inserted before Function A at the start, and on the whole, rather than Function D (departure) being followed by Function E (the first function of the donor vs. the hero's reactions) as in the case of Group A (with the exception of P133-2), it is followed by a function such as Function K (the initial misfortune or lack is liquidated), Function G (spatial transference between two kingdoms, guidance), or Function F (provision or receipt of magical agent). From these patterns the following rough description for the motif for the start sections of Group B is obtained.

[EF]ACD[KGF] . . .

The previous discussion was in regard to the parts at the start of the sequences. It is also probably possible to convert the common elements Function C and Function D as fixed chains to specific codes and increase the alignment sensitivity by performing alignment again, and to extract other specific parts and perform alignment to perform local comparisons.

TREEVIEW can be used to output the alignment results in Radical and Phylogram format and facilitate observation. Figures 6 and 7 show the Radicals and Phylograms based on the alignment output for Group A, while Figures 8 and 9 show the Radicals and Phylograms based on the alignment output for Group B.

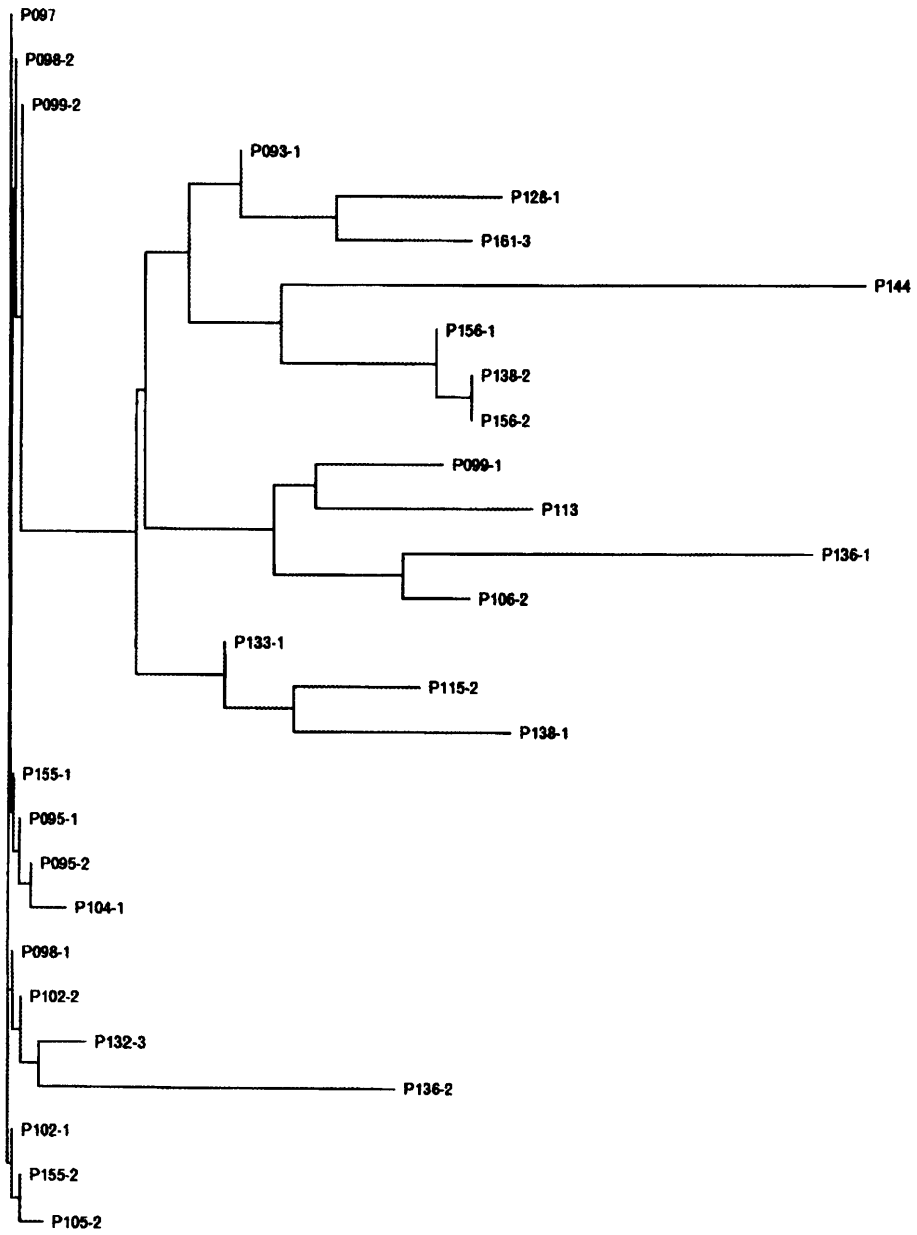

Fig. 6 Phylogram of Group A

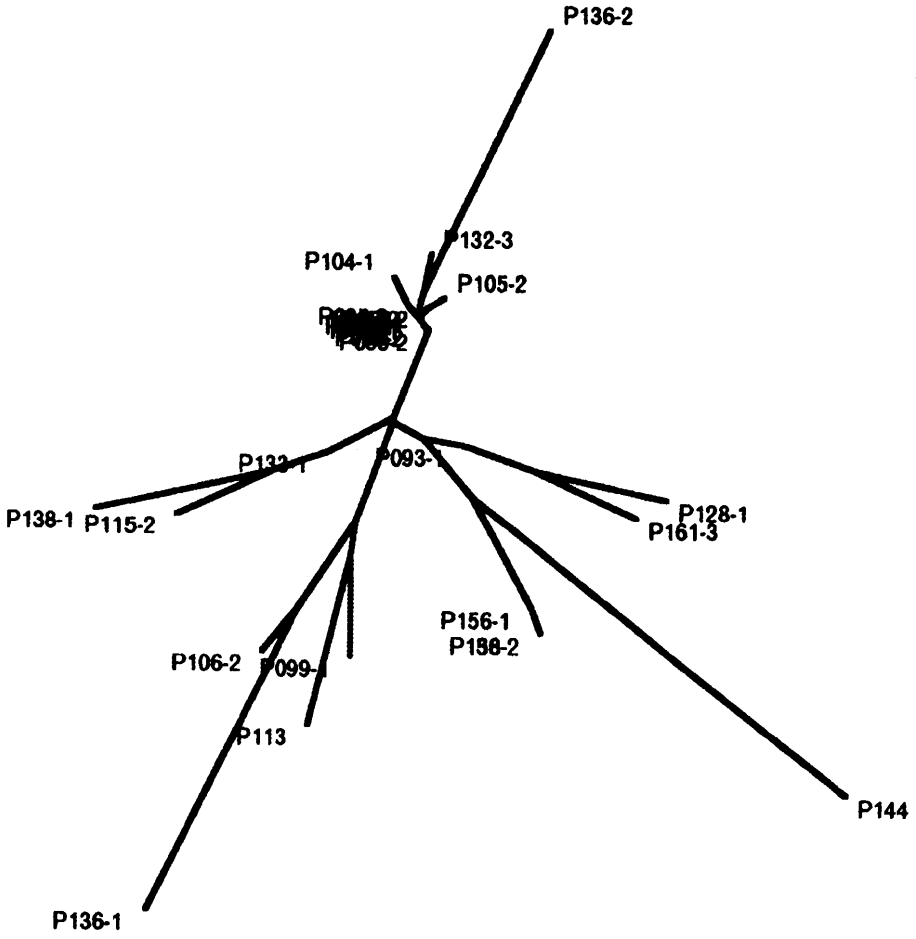

Fig. 7 Radical of Group A

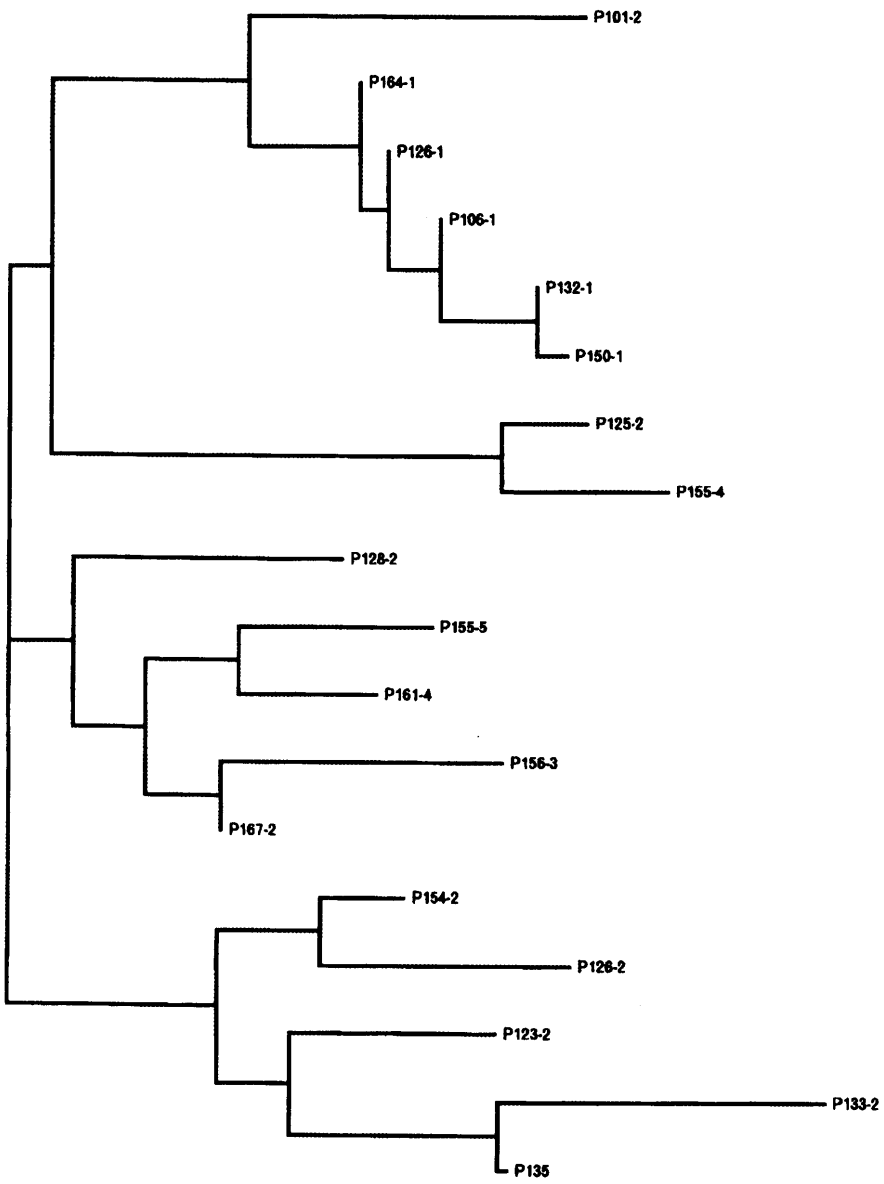

Fig. 8 Phylogram of Group B

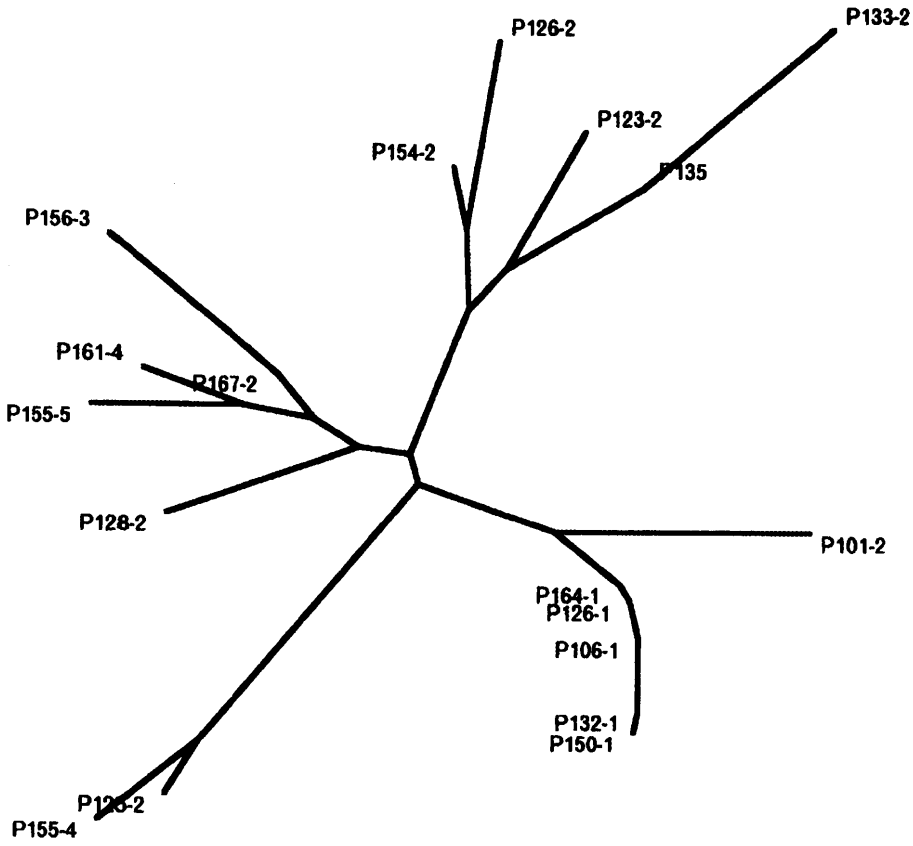

Fig. 9 Radical of Group B

With the Phylograms, the high-similarity sequences are concentrated on the left side of the tree, and the low-similarity sequences are separated and placed on the right side of the tree. On the other hand, with the Radicals, the distance from the central point exhibits an inversely proportional relationship with the similarity (in other words, when the similarity is high, the mutual distance is short). These figures indicate clearly that all sequences determined to be 'exceptional' in the alignment output are far away. For example, P144, P136-1, P136-2, and P138-1 in Figures 6 and 7, and P101-2, P133-2, and P155-4 in Figures 8 and 9, are separated from the other sequences. It is also possible to observe that the grouping of sequences was performed using function types that followed Function F in the case of the Group A alignment output, and function types that followed Function D in the case of the Group B alignment output. Combining this with the above alignment output seems to provide suggestions for indexing when classifying sequences.

With regard to determining whether the analysis results are significant or not, it is clear that lining up the text analysis of individual tales and treating the sequences of elements that form the high-similarity sequences that require judgment as motifs, then comparing these to low-similarity sequences and identifying the insertions and deletions of specific functions provides a starting point for more detailed analysis.

CONCLUSION

Through the alignment and its analysis attempted in this paper, I have been able to clearly establish that the analysis methods used for bioinformatics are an effective means for analyzing folktales from several perspectives, and visual presentation of the phylogenetic relationship of sequences enables a lot of information to be garnered. To apply the analysis methods used for bioinformatics more effectively to the analysis of folktales in the future, it will be necessary to investigate both technical problems from an informatics perspective, and narratology problems relating to how to best describe the sequence structure of folktales as data.

An example that relates to the first issue involves the speed and sensitivity limitations of CLUSTAL W when it applied as is to analysis of the structure of the folktale, and this has been clearly established in this paper. As CLUSTAL W was developed for operation on relatively small computer systems, it is effective when used for alignment of sequences that are similar to a certain degree, but to process sequences such as those for folktale structure used in this paper, additional operations are required to improve the sensitivity. Among the alignment programs used for bioinformatics, there are examples that employ different algorithms to CLUSTAL W, and other analysis tools and new software are continually being developed.[15] I am testing these, as far as possible, while waiting for more sensitive tools to be developed. In addition, I am also looking at tools that do not run on the system currently being used (e.g. tools developed at the ICOT[16] etc.) and am investigating the possibility of transferring them to high-performance computer systems and mainframes. For this purpose I plan to enlist the cooperation of specialists in the fields of informatics and biology.

With regard to the second problem, there is a need to construct a model that optimally describes the sequences used in the structure of folktales. As I stated at the start of this paper, other papers using description models from the results of case grammar, semantic structure and artificial intelligence are currently being prepared.

Finally, I would like to add a few words about unification of analysis methods for folktale structure. Research into analysis methods for folktales and similar texts can be divided into two types: research that focuses on syntagmatic structure carried out by Propp and others, and research that emphasizes the paradigmatic structure such as the work of Lévi-Strauss[17] (there are also analysis methods that handle both these structures[18]). Methodologically speaking, this paper has focused on analysis that uses the syntagmatic structure of the folktale, but by combining this with paradigmatic structure analysis methods, I believe that more fruitful folktale

analysis will be possible. In my past numerical research into paradigmatic structure,[19] I obtained some results with regard to the conversion procedure stage for converting folkloristic text into data for processing (individuals [= tale] × variables), but the problem of the arbitrariness of the settings of variables arose. In that research, by handling Propp-like functions[constants] and the variables that form the functions in the same dimension, I was able to obtain the trend characteristics of the genre being analyzed from the co-occurrence pattern of the specific constants and specific variables, but this was nothing more than a rough result.

The objectives of the unified description model that I am currently hypothesizing for the structure of the folktale are to reevaluate the syntagmatic structure as a paradigmatic structure with the entire potential insertion/deletion pattern of the manifestable elements included, and conversely, to describe the variables handled as the configuration in the text as the syntagmatic structure in their order of appearance for the text surface structure. By doing this, it will be possible to simultaneously analyze both the syntagmatic structure and paradigmatic structure. In other words, it will realize an analysis method that unifies morphology and transformation. In addition, it will probably become possible for conventional narratology to unify the two different directions of deep structure [content] and surface structure [expression] that have become the subject of analysis.

NOTES

1) Oda 1997.
2) Seki 1981 (1955), p.60. Boogaard 1984, p.667.
3) Propp 1968 (1928).
4) Greimas 1966.
5) Beatie 1979, p.194.
6) Example from Uchiyama, I. *et al.* (Mitaku, S. and M. Kanehisa (eds.) 1995).
7) Thompson, J. D., Higgins, D. G. and T. J. Gibson 1994.
The program was downloaded from the European Bioinformatics Institute site (http://www.ebi.ac.uk/)
8) Page 1996.
The program was downloaded from the Ribosomal Database Project FTP Server (University of Illinois at Urbana-Champaign).
9) Propp 1968 (1928), p.21.
10) ibid., pp.135-143.
11) When alignment was actually performed on these sequences, it was discovered that the distances between sequences were abnormal.
12) The 31 functions established by Propp were reduced to 11 function pairs (22 functions) by Greimas. Also, the seven functions (β, γ, δ, ε, ζ, η, θ) for the preparatory part of the tale of Propp's Schemes were omitted. In addition, Greimas eliminated Propp's Function L (unfounded claims), and as a result of this reduction, when Propp's Schemes are converted to the Greimas model, 16 functions result.

13) For proteins, CLUSTAL W recognizes the following 20 amino acids. A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y.
14) This is called the FASTA format (Pearson and Lipman).
15) http://www.ebi.ac.uk/biocat/Alignment_Search_software.html
16) Institute for New Generation Computer Technology
http://www.icot.or.jp/AITEC/3/IFS/list/catalogue-J.html#experimental
17) Lévi-Strauss 1958.
18) Barthes 1977 (1966).
19) Oda *et al.* 1984 and Oda 1984.

REFERENCES

Barthes, R.
1977 (1966) Introduction à l'analyse structurale des récits. *Poétique du Récit*, pp.7-57. Paris: Éditions du Seuil.

Beatie, B.A.
1979 Measurement and the Study of Literature. *Computers and the Humanities* 13, pp.185-194.

Boogaard, N. van den
1984 La définition du fabliau dans les grands recueils. *Epopée, animale, fable, fabliau. Cahiers d'Études Médievales* 2-3, pp.657-668. Paris: P.U.F.

Greimas, A.-J.
1966 *Sémantique Structurale*. Paris: Larousse.

Lévi-Strauss, C.
1958 *Anthropoplogie Structurale*. Paris: Plon.

Mitaku, S. and M. Kanehisa (eds.)
1995 *Human Genome Program and Knowledge Information Processing*. Tokyo: Baifukan. [in Japanese]

Oda, J. *et al.*
1984 Methods of Bibliographic Information Analysis and Structure of Literary Text. *Proceedings of International Computer Symposium 1984 (ICS '84)*. vol. II, pp.908-915. Tamkang Univ.

Oda, J.
1984 Structure du narration du fabliau: Essai d'un reclassement des pièces par Cluster-Analysis. *Mathematical Linguistics* 14-7, pp.281-303.
1997 Description of Structure of the Folktale I: From Bioinformatics to Text Theory. In J. Oda (ed.), *Genesis of Narrative*, pp.67-92. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies. [in Japanese]

Page, R. D. M.
1996 TREEVIEW: An Application to Display Phylogenetic Trees on Personal Computers. *Computer Applications in the Biosciences* 12, pp.357-358.

Propp, V.
1928 *Morfológija skázki*. Leningrad. (English edition translated by Laurence Scott, 1968, *Morphology of the Folktale*. Austin: University of Texas Press)

Seki, K.
1981 (1955) *Structure of the Folktale*. Tokyo: Dohosha. [in Japanese]

Thompson, J.D., Higgins, D.G. and T.J. Gibson
1994 CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Positions-specific Gap Penalties and Weight Matrix Choice. *Nucleic Acids Research* 22, pp.4673-4680.