사상체질의학회지
J. of Sasang Const. Med.
Vol. 13. No 1. 2001

생물정보학을 이용한 체질정보의 정량에 관한 이론적 고찰

지상은* · 한성규* · 최선미**

Abstract

Theoretical study on the quantification of constitutional information using bioinformatics

Chi Sang-eun* · Han Sung-kyu* · Choi Sun-mi**

* Genopia (Dong-ui Clinic of oriental medicine)
** Korea Institute of Oriental Medicine

Purpose

This study was carried out to apply the knowledge of bioinformatics to the quantification of constitutional information.

Methods

To objectify consititutional knowledge, several uselful methods including Bayesian estimate, position specific score matrix, entropy, phylogenetic tree, simulated annealing were discussed.

Results and Conclusion

It is obvious that bioinformatic methods can be the most important tool for the objectification of constitutional medicine.

Key Words : constitution, bioinformatics, quantification

Ⅰ. 緖 論

생물정보학(bioinformatics)은 "생명현상 연구에 필요한 다양한 전산학/통계학/수학적인 것들"에 대한 학문으로 표현될 수 있다[1].

생물정보학은 휴먼게놈프로젝트의 수행 과정에서 그 중요성이 부각되었으며, 휴먼게놈프로젝트 결과를 해석하고 가공하는데 핵심적인 역할을 하고 있는데, genomics적인 실험에 수반되는 data의 양과 처리 방식이 기존의 분자생물학적인 data의 처리방식과는 차원이 틀린 복잡성을 가지기 때문에, 결국 통계적인 처리나 모델링 방법을 사용할 수 밖에 없게 된다. 생물정보학의 핵심은 단편적이고, 단면적인 정보들을 가공하여 보다 높은 차원의 실용적인 정보나 모델로 구축하는 것이다.

인체에서 나타나는 정보는 복잡한 현상의 한 단

* (주)제노피어(동의한의원) ** 한국한의학연구원
교신저자 : 지상은 주소)경기도 고양시 덕양구 화정동 984-3 세일빌딩 303호 전화)031-973-1010 E-mail)mtmind@hitel.net

면으로, 이러한 정보를 통해 全一體의 상태를 추론하는 것이 동양의학의 해석론인 藏象學의 원리이며, 形象的 체질의학의 원리라 할 수 있다. 그러나 생체에서 나타나는 방대하고 복잡한 정보들을 구체적으로 처리, 가공하여 모델화할 수 있는 방법론은 최근까지도 존재하지 않았기 때문에, 동양의학에서는 이러한 정보들을 陰陽, 五行, 四象과 같은 類聚 방법을 통해서 범주화시켜서 파악하였다. 현재의 생물정보학은 컴퓨터의 발달과 함께 바로 이러한 복잡한 data를 다룰 수 있는 방향으로 영역을 확대해 나가고 있으며, 따라서 동양의학의 약점을 보완하는 유용한 도구로 활용될 수 있을 것으로 사료된다.

이에 저자는 생물정보학에서 사용되는 몇가지 방법론을 원용하여 체질의학적인 정보를 정량화하고 객관화하는 방법에 대하여 고찰하였다.

II. 본 론

1. 체질의 판단 절차

체질을 판단한다는 것은 일반적으로 다음과 같은 절차(procedure)로 설명할 수 있다.

(1) 모델선택 : 체질에 대한 모델을 선택한다.

(2) 정보수집 : 환자로부터 모델에 따른 체질 정보를 수집한다.

(3) 정보가공 : 체질정보를 가공한다.

(4) 정렬 (alignment) : 모델과 환자로부터의 체질 정보를 정렬 (alignment)을 통해 optimal alignment 또는 near optimal alignment를 찾아낸다.

(5) 판단 : 찾아낸 alignment가 일정한 역치 (threshold) 이상의 score일 경우 accept하고, 만약 역치 이상의 score가 나오지 않을 경우 reject한 후 (2), (3), (4)번 또는 (1)번 중 하나로 회귀한다.

이러한 절차는 다음과 같은 nondeterministic finite state automata로 표현할 수 있다. (Fig. 1.)

여기서 다음과 같은 몇가지 부분이 지적될 필요가 있다.

1. (1)~(5)번의 과정 전체에서 분명한 기준을 정하지 않는다면 정보의 변형 또는 왜곡이 일어날 가능성이 항상 존재한다.

2. (2)번에서 체질정보의 수집 과정에서는 의사로부터 기인한 정보 변형 뿐 아니라 환자로부터의 정보 변형이 일어날 가능성도 존재한다. 환자가 의도적으로 정보를 왜곡, 또는 숨기거나 충분히 표현하지 못하는 경우가 있을 수 있다. 이러한 관점에 따라 체질정보를 3가지로 나눌 수 있는데 첫째는 완벽히 객관적인 정보로 수치화된 것이다. 키, 몸무게 등과 같은 것이다. 둘째는 환자가 표현한 그대로의 정보이다. 자각증상과 같은 경우이다. 셋째는 환자가 표현하거나, 환자로부터의 정보를 의사가 취사선택하는 것이다. 이 세 번째 과정에서는 필연적으로 (3)번에 해당되는 정보의 가공이 일어나게 된다.

3. (3)번 체질정보의 가공은 (2)번에서 수집한 정보들을 취사선택하고 연관을 시킬 것인가의 문제로, 어떤 체질 모델을 선택했는가에 따라 주로 결정되는 부분이다. 이 부분에서 의사간의 편차가 심하게 나타날 수 있다.

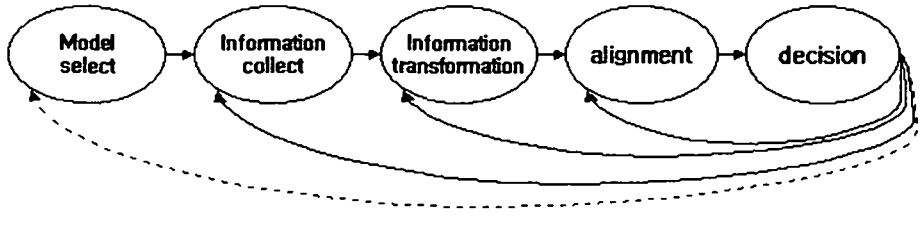

Fig. 1. Procedure of discrimination of constitution was displayed in the form of nondeterministic finite state automata.

4. 정렬 (alignment)은 의사의 모델과 체질 정보를 match시키고 match와 mismatch에 따라 점수화 (scoring) 하는 부분이다.

여기서 한가지 중요한 점은 이 부분을 어떻게 정의하는가에 따라서 정량화할 수 있는 수학적 모델을 구성할 수 있는가 없는가의 차이점이 나타나게 된다는 것이다.

이는 촘스키의 생성문법론(generative grammar, transfomational grammar)의 주된 내용으로, 그는 "어떠한 언어(language)가 특정한 문장(sentence)을 포함하는가?"의 질문을 "어떠한 문법(grammar)이 이러한 특정 문장(sentence)을 생성할 수 있는가?"의 질문으로 대치하였다[2]. 처음의 질문은 가능한 문장의 수는 무한하기 때문에 수학적으로 polynomial한 time 내에 풀릴 수 없는 'intractable'한 NP problem (nondeterministic polynomial problem)인 반면, 후자의 질문은 문법을 어떻게 모델하는가에 따라 polynomial한 time에 실제적으로 풀릴 수 있는 문제가 된다. 여기서 설명한 (4), (5)번은 기본적으로 이러한 생성 모델에 근거하여 구성된 것이다. 이러한 생성 모델의 예는 biological sequence analysis에 사용되는 stochastic context free grammar와 hidden Markov model에서 볼 수 있다.

5. (5)번에서 자신의 모델과 환자로부터의 체질정보 사이의 alignment의 score가 기준 이상으로 나오지 않을 경우 (2), (3), (4)번으로의 회귀는 쉽게 일어날 수 있지만, 한 개인에서 (1)번으로의 회귀는 확률적으로 더 작다는 점이다. 왜냐하면 체질감별의 모델이 대부분 어떠한 분명한 정의 (definition)에 의해서 이루어진 모델이 아니라, 뇌의 신경망 (neural network)에서 학습 (training)된 경험적 (heuristic) 모델인 경우가 많기 때문에, 개인에게 있어 모델의 변경 (change)은 어려운 경우가 많다. 다만 체질모델에서 證 모델로의 변경과 같이 다른 차원으로의 변경은 흔히 일어나는 일이라고 볼 수 있다. 체질이 애매한 경우 기존의 변증모델에 근거하여 진단하는 것은 일반적인 일이다.

2. 모수의 추정 (parameter estimation)

체질을 판단한다는 것은 현실적으로는 확률적인 문제로 귀결된다.

확률적 모델이 선택된 후에는 모집단의 특성인 모수 (parameter)가 실제적인 체질감별의 결과로 나온 data로부터 추정되어져야 한다.

이러한 모수를 추정하는 방법은 maximum likelihood estimation과 Bayesian estimation이 있다[2].

Maximum likelihood estimation은 고전적인 방법으로 실제 data에서 나온 그대로 설정하는 방법이다. 이것은 합리적인 방법이지만 data의 수가 적을 때에는 문제가 될 수 있으며, 체질의 감별에 있어서는 아무리 동일한 조건으로 체질을 감별하였다 하더라도 서로 다른 사람의 감별 사이에는 차이가 있을 수 있기 때문에, 이를 보완하기 위하여 다음과 같은 Bayesian estimation이 더 적절하다고 보여진다.

어떤 지표를 가지고 어느 체질에 해당할 것인지에 대한 확률을 계산하는 것은 Bayes' theorem[3]에 의해

$$p(\text{체질} \mid \text{지표}) = \frac{p(\text{지표} \mid \text{체질})p(\text{체질})}{p(\text{지표})} = \frac{p(\text{지표} \mid \text{체질})p(\text{체질})}{\sum_{\text{체질}} p(\text{지표} \mid \text{체질})p(\text{체질})}$$

으로 나타낼 수 있다.

여기서 p(체질)은 prior probability이고, 우리가 알고자 하는 것은 posterior probability인 p(체질 | 지표)이다.

Prior probability p(체질)는 『東醫壽世保元』「四象人辨證論」에 나타난 체질별 분포가 東武 李濟馬 당시의 지역적 특성과 무관한 보편적인 것이고, 현재까지 분포의 변화가 없었다고 가정한다면 p(태음) = 0.5, p(소양) = 0.3, p(소음) = 0.2 이다. 이것은 보다 현실적인 data에 의해서 대치될 수 있을 것이다.

그리고 p(지표 | 체질)은 지표에 따라 틀리며, 실제적인 data로부터 알 수 있는 부분이다.

이와 같이 우리가 각 체질에 대한 분포와 그 체질에서 지표의 확률을 알 수 있다면 그 지표를 가지고 특정 체질에 해당하는 확률을 추정해 낼 수 있으며, 이를 점수화하여 계량화할 수 있다.

특히 Bayesian estimation이 반드시 적용되어야 하는 경우는 실제 data에서 태양인의 확률이 0인 경우

정함에 따라 수학적인 여러 가지 문제점을 극복할 수 있다.

3. Position specific weight matrix를 사용한 p(체질 | 지표)의 산정

전술한 지표에 대한 확률은 그 자체를 점수로 쓸 수 있다.

그러나 이때 문제가 되는 것은 지표가 하나가 아닐 경우 점수는 확률의 곱으로 표현되기 때문에 컴퓨터에서의 계산상 underflow를 일으키게 된다. 따라서 이를 극복하는 유용한 방법은 확률에 log를 취하는 것이다. 이때의 문제점은 확률이 0인 경우에는 log를 취할 수 없다는 것인데, Bayesian estimation에서는 pseudocount를 넣어줌으로써 극복될 수 있다.

이러한 점수를 구하기 위해서는 미리 확률적으로 계산된 score matrix를 활용하는 것이 유용하다.

특정 체질, 어떤 지표에서 나타나는 반응을 몇가지로 구분한 후 해당되는 count의 확률을 모든 체질에서 나타나는 frequency로 나눈 log odd ratio의 matrix 가 position specific weight matrix가 된다.

$$Score = \sum_{\text{반응}i} \frac{\text{특정 체질에서 나타나는 반응 } i\text{의 확률}}{\text{전체 체질에서 나타나는 반응 } i\text{의 확률}}$$

예를 들어, 소음인 20명에서 3개의 지표들에 대한 다음과 같은 count가 주어졌을 때, position specific weight matrix를 구할 수 있다(Table 1.).

Table 1. Arbitrary Data for Calculating Position Specific Weight Matrix

Soumin					Total constitutional group
Reaction	Index1	Index2	Index3	Frequency	Random Frequency
A	5(+1)	17(+1)	10(+1)	35 / 69	0.1
B	10(+1)	1(+1)	10(+1)	24 / 69	0.5
C	5(+1)	2(+1)	0(+1)	10 / 69	0.4
Total	23	23	23	69 / 69	1.0

+1 counts in parentheses are pseudocounts.

Bayesian estimation을 위해 Laplace rule에 의해서 모든 count에 각각 1의 pseudocount를 가하였는데, 이는 지표 3의 반응 C가 0으로 나온 문제점을 해결하기 위해서이다. 그런데 이 pseudocount를 조절함으로써 현재 이미 알려진 지식을 활용할 수 있다는 것이 Bayesian estimation의 장점으로, 만약 지표 2에서 소음인은 반응 A 가 이론적으로 더 기대된다면 A에 보다 높은 pseudocount를 주고 그만큼을 B와 C에서 뺄 수 있는 것이다.

이를 편의상 base가 10인 log를 취하여 계산하면 다음과 같다.

Table 2. Position Specific Weight Matrix

	Index1	Index2	Index3
A	0.77	1.25	1.04
B	0.34	-0.39	0.34
C	0.17	-0.12	-0.60

따라서 소음인의 경우 지표 1에서 A가 나오는 score 는 0.77이며, 이는 p(체질 | 지표)에 log를 취한 값이다. 이 score에 대해서는 확률적인 significance를 구하는 것이 가능하다.

어떤 사람이 지표 1에서 A, 지표 2에서 C, 지표 3에서 B가 나왔을 경우, 이 사람이 소음인일 score는 0.77−0.12 + 0.34 = 0.99이다.

4. Entropy 개념을 이용한 체질 정보의 정량

Entropy의 개념은 정보이론 (information theroy)에서 제출된 것으로, 엔트로피란 결과의 평균적인 불확실성의 측정이다[2].

임의의 (random) 변수 X에 대한 K events $x_1, \ldots, x_K$ 의 이산 (discrete) 확률 $P(x_i)$가 주어질 때,

Shannon의 entropy[2]는 다음과 같이 정의된다.

$$H(X) = -\sum_i P(x_i)\log P(x_i)$$

$$(P(x_i)\log P(x_i) = 0, \text{ if } P(x_i) = 0)$$

일반적으로 log는 base를 2로 하게 되며, 이때 엔트로피의 단위는 'bit'가 된다.

트로피의 단위는 'bit'가 된다.

엔트로피는 모든 $P(x_i)=\frac{1}{K}$ 일 때 최대화되며, 한 k에 대해서 $P(x_k)=1$이고, 다른 $P(x_i)=0$일 때 엔트로피는 0이 된다. 보통 엔트로피는 정보(information)과 동일시되는데, 보다 random할수록(엔트로피가 높을수록) 정보가 더 많아지게 된다.

다른 중요한 엔트로피의 측정방법의 하나로 'mutual information'[2]이 있으며 다음과 같이 정의된다.

$$M(X;Y)=\sum_{i,j}P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)}$$

Mutual information은 X와 Y가 항상 같이 변할 때(covary) 최대화된다.

체질정보들은 서로 다른 엔트로피를 가질 것으로 생각된다. 각 체질별로는 호발되는 증상과 증상의 변화 폭이 다를 것이다. 예를 들어 소화불량, 식체가 다른 체질보다 소음인에서 더 많이 나타난다고 볼 때, 소음인에게서는 소화기 증상의 엔트로피가 낮을 것이며, 다른 체질에서는 엔트로피가 높을 것으로 기대할 수 있다. 이는 소음인은 주로 소화기능이 약화되는 증상이 호발하지만, 다른 체질에서는 소화기능이 약한 사람부터 소화기능이 강한 사람까지 보다 다양할 것으로 생각되기 때문이다. 만약 소음인이 전적으로 소화기능이 약하고, 그 반대인 소양인은 반대로 전적으로 소화기능이 강하며, 그 변화폭이 동일하다면 엔트로피는 같을 것이다.

이러한 방법으로 다양한 증상에 대하여 많은 량의 data 로부터 추정한다면 의미있는 결과를 얻을 수 있을 것으로 생각된다.

또한 『東醫壽世保元』 「太陽人 內觸小腸病論」에 근거한다면 구토의 증상이 있을 때, 소음인과 소양인은 한열의 증상이 covariation이 높아서 높은 mutual information을 나타낼 것으로 생각된다. 이러한 covariation은 체질별로만 존재하는 것이 아니라 병증별로도 존재할 수 있을 것이다.

事心身物의 체계는 자유도에 따른 체계로 가정할 수 있으며[4], 이러한 관점에서는 엔트로피의 개념과 밀접한 관련이 있다고 볼 수 있다. 자유도가 가장 높은 事의 차원은 엔트로피가 높기 때문에 다양한 정보가 나타날 수 있다. 이에 비하여 자유도가 가장 낮은 物의 차원은 엔트로피가 낮기 때문에 일관된 정보가 나타날 것이다. 엔트로피의 개념은 정보들을 정량화하는 유용한 도구로 다양한 분야에 일관되게 적용할 수 있다는 것이 큰 장점으로 생각된다.

5. Phylogenetic tree를 이용한 체질감별방법의 모델링

체질의학의 인식적인 방법은 어떠한 계층을 이루고 있다. 그 계층은 사상철학적으로는 事心身物이라고 할 수 있으며, 다양성과 자유도를 기준으로 논할 때는 전술한 엔트로피의 개념으로 설명되어질 수도 있을 것이다.

『東醫壽世保元』에 나타난 체질 감별의 원칙들을 모델화하기 위해서는 각 문장과 단락에서 나타난 병증 또는 체질감별의 원칙들을 하나의 규칙으로 간주하고, 보다 일반적인(general) 규칙들을 ancestor로 하는 phylogenetic tree를 구성할 수 있을 것이다. 이러한 보다 일반적인 원칙들은 주로 체질별 保命之主에 대한 것일 확률이 높을 것이다.

이러한 tree의 특성은 많은 환자들의 자료를 바탕으로 re-estimation 되어질 수 있을 것이다.

6. local optima

Local optima란 global optima와 상대되는 용어로 부분적인 최적치를 뜻하며, 근사치를 추정하는 알고리즘이 빠지기 쉬운 오류 중의 하나이다.

체질감별의 문제에 있어서는 더욱 이런 현상이 많을 것으로 생각된다.

이러한 문제점을 극복하기 위해서 'simulated annealing'과 같은 알고리즘이 사용되고 있다. Krogh 등은 이를 바탕으로 data에서 나온 count에 noise를 더함으로써 유사한 효과를 얻은바 있으며[2], 이 방법은 체질감별 알고리즘에 그대로 적용되어질 수 있다.

실제적으로 체질별로 존재하는 가능한 모든 병증에 대하여 환자의 data와 model 간의 alignment를 통해서 optimal alignment를 찾는 것은 가능하지 않을 수도 있다. 이러한 경우에는 time이나 memory의 complexity를 줄이는 알고리즘을 사용하게 되는데,

운 약점이 있다.

이는 비유를 하자면, 소양인적인 특징과 태음인적인 특징을 동시에 가진 어떤 환자에게 열다한소탕이 어느 정도 좋은 반응을 보이면서 특별한 부작용이 없을 경우, 이 사람이 재차 내원했을 때 소양인 처방을 쓰는데는 어려움이 있는 것과 유사하다. 하지만 그것이 가장 최선의 방법이었는지는 알 수가 없게 된다. 이러한 경우, 몇 번에 한번 꼴로 후자를 선택한다면 local optima에서 벗어날 확률이 발생하게 된다.

이러한 문제는 전체적인 의료의 차원에서도 생각할 수 있다. 현재 체질감별의 기법들, 즉 model들은 여러 가지 종류가 존재하며, 이들이 공존하는 것이 현실이다. 한국의 제도상 환자는 진료기관을 선택할 수 있기 때문에, 자신의 체질 혹은 다른 어떤 조건을 만족시키는 진료기관을 찾아서 계속 옮겨다닐 수 있다. 이러한 유동이 매우 활발히 일어난다고 가정한다면, 각 진료기관마다 그 진료기관의 조건에 맞는 환자들만이 남는 수렴(convergence) 현상이 발생한다. 그 조건은 체질감별의 방법일 수도 있고 다른 어떤 것, 또는 복합적일 수도 있으나 일단 체질감별방법이라고 가정한다면, 의사는 자신의 체질감별방법에 대한 local optima에 빠질 가능성이 높다. 자신의 감별방법에 맞지 않는 환자는 다른 진료기관으로 가고, 자신의 감별방법에 맞는 환자들만이 계속 래원할 것이기 때문이다.

물론 이러한 가정은 가정일 뿐이며, 현실적이고 구체적인 data로 다룬다는 것이 불가능한 성질의 것일 가능성이 높다. 하지만 local optima를 제거할 수 있는 장치나 제도가 한방의료 시스템 내에 존재하는가 존재하지 않는가의 문제는 실제적으로 다루어질 수 있는 문제일 것이다.

7. 생성문법을 이용한 체질 정보의 parsing

이러한 정량화를 이루기 위해서는 체질정보의 입력 단계에서 디지털화되는 것이 바람직하며 필연적으로 전자챠트가 필요하게 된다.

가장 단순한 방법으로는 필요한 정보들로 구성된 관계형 database를 설계하는 것이다. 그러나 이러한 방법은 사용자 친화적이지 않고 비효율적이며, 확장성에 한계를 가지는 것으로 판단된다.

따라서 text형식으로 사용자가 입력할 때 이를 parsing하여 적절하게 변환해주는 프로그램이 필요하다. 하지만 이때 사용하는 언어가 자연어(natural language)라면 체질감별에 대한 연구보다 자연어 인지에 대해서 훨씬 많은 연구를 해야 할 것이다. 따라서 이러한 언어의 문법은 촘스키의 hierarchy 중 regular grammar의 수준에서 정의되는 것이 바람직할 것이다.

Ⅲ. 고찰 및 결론

본 연구에서는 생물정보학적인 방법론을 통하여 체질정보를 어떻게 정량할 것인가에 대하여 이론적으로 고찰하였다. 그러나 언급된 부분들은 생물정보학의 극히 지엽적인 일부분에 불과하며, 동양의학에 접목될 수 있는 수많은 가능성 중의 하나라고 할 수 있을 것이다.

충분한 data만 확보된다면 생물정보학은 동양의학에서의 많은 경험적 정보들을 객관화하는 방법으로 사용될 수 있을 것이다. 예를 들어, 'data driven discovery'라고 불리워지는 방법론은 각 체질 당 10만 명 이상의 체질이 감별된 사람들을 확보한다면, 몇 만개 수준의 많은 유전자들에 대한 cDNA microarray를 통해서 이들의 data를 분석하여, 이를 바탕으로 체질을 알지 못하는 어떤 사람이 무슨 체질에 속할지를 확률적으로 진단해 낼 수 있다는 것이다. 이러한 결과는 체질이 무엇인가에 대한 답을 주지는 못하지만 상당한 확률을 가지고 체질을 진단해 낼 수 있을 것이다. 체질 뿐 아니라 질병이나 수명 등 어떠한 정보라도 유전자에 coding되어 있는 정보라면 이러한 방법론으로 예측해 낼 수 있을 것을 것으로 생각된다.

이와 같이 생물정보학적인 지식들은 현재까지 효율적으로 기술하지 못했던 생체의 복잡계에 대한 해석을 가능하게 하고 있으며, 이는 동양의학의 객관화에 무엇보다도 절실했던 방법론이라 할 수 있다. 현재 시기는 생물정보학의 태동기라 할 수 있기 때문에 비교적 경쟁력을 키울 수 있는 시기이나, 몇 년 후에는 이러한 지식도 선진국과의 격차가 도저히 따라갈 수 없을 정도로 벌어질 것이 자명하기 때문에, 이러한 학문에 대한 집중적인 연구와 투자

가 시급히 필요할 것으로 생각된다.

參 考 文 獻

1. 원세연. 생물정보연구소, 과학기술정책, 2000년 9월호.
2. R. Durbin, S. Eddy, A. Krogh, G. Mitchison. Biological Sequence Analysis Cambridge university press. 2000: 5-10, 154-155, 234, 305-310.
3. Kenneth H. Rosen. Discrete Mathematics and It's Applications. McGRAW-HILL. 1999: 305.
4. 지상은, 조황성. 사상체질의학의 진화론적 고찰. 한국한의학연구원 논문집 제3권 제1호. 1997; 3(1) :105-118.