УДК 61:004.8

Д.А. БОЙКО, студент, НТУ "ХПИ" (г. Харьков),
О.В. ВАСИЛЬЕВА, мл. науч. сотрудник Украинского института клинической генетики, ХНМУ (г. Харьков),
Д.А. ГАЛКИН, студент, НТУ "ХПИ" (г. Харьков),
Ю.Б. ГРЕЧАНИНА, канд. мед. наук, доц., зам. директора по лечебной работе ХСМГЦ (г. Харьков),
А.И. ПОВОРОЗНЮК, канд. техн. наук, доц. НТУ "ХПИ" (г. Харьков),
А.Е. ФИЛАТОВА**,** канд. техн. наук, доц. НТУ "ХПИ" (г. Харьков)

СОЗДАНИЕ ИНФОРМАЦИОННОЙ СТРУКТУРЫ БАЗЫ ДАННЫХ КОМПЬЮТЕРНОЙ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ ДИАГНОСТИКИ МИТОХОНДРИАЛЬНЫХ ЗАБОЛЕВАНИЙ

В работе рассматривается задача формализации исходных данных при диагностике митохондриальных заболеваний. В качестве исходных признаков были рассмотрены фенотип, биохимические исследования крови и мочи, а так же заболевания в родословной и сопутствующие диагнозы. В результате разработаны справочники для ведения базы данных (БД). Работа ведется совместно с Харьковским специализированным медико-генетическим центром.

Ключевые слова: формализация исходных данных, митохондриальные заболевания, фенотип, биохимические исследования крови и мочи, база данных.

Постановка проблемы. Митохондриальные заболевания (МЗ) – это группа наследственных заболеваний, связанных с дефектами в функционировании митохондрий, приводящих к нарушениям энергетических функций в клетках эукариотов. Для постановки диагноза МЗ важен комплексный генеалогический, клинический, биохимический, морфологический и молекулярный анализ. Создание компьютерной системы поддержки принятия решений (КСППР) для диагностики МЗ является актуальной научно-технической проблемой. Одним из основных этапов создания КСППР является разработка специализированной базы данных (БД), структура которой позволит легко добавлять не только количество пациентов, но и менять множество признаков, необходимых для диагностики.

Анализ литературы. На сегодняшний день имеется достаточно четкое представление о причинах МЗ [1, 2]. Они обусловлены генетическими, структурными, биохимическими дефектами митохондрий и нарушением тканевого дыхания. Генетические дефекты дыхательной цепи и возникающая в результате этого недостаточность аденозинтрифосфатной кислоты (АТФ) нарушают многочисленные функции клеток, что особенно проявляется в высокоэнергетических органах [3]. Хотя наибольшей потребностью в митохондриальной энергии обладают нейроны, скелетная мускулатура,

сердечная мышца, клетки костного мозга и эндокринные железы, ее хронический недостаток может привести к патологическим изменениям практически в любом органе [2, 4]. Поэтому для диагностики МЗ важно комплексное изучение клинико-генетических характеристик больных.

Для правильной постановки диагноза митохондриальной болезни (митохондриопатии) необходимо применять как классические методы исследования – соматогенетическое исследование с синдромологическим анализом и клинико-генеалогический анализ, так и современные методы биохимической и молекулярной диагностики [5, 6]. В связи с множеством параметров, которые оцениваются при подозрении на МЗ, целесообразным является создание специализированной компьютерной БД, а также разработка современных методов статистического анализа, адаптированных к клинической практике [7 – 10]. На рис. 1 представлена общая схема анализа экспериментальных данных при проектировании КСППР в медицине.

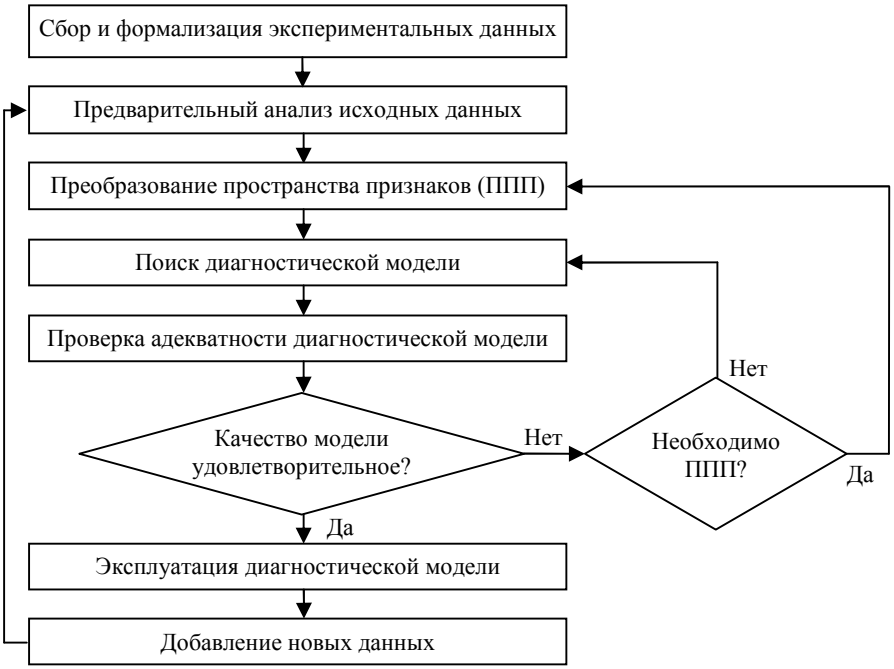

Рис. 1. Общая схема анализа экспериментальных данных

Данная работа направлена на реализацию этапов сбора, формализации и предварительного анализа экспериментальных данных при диагностике МЗ.

Целью данной статьи является анализ исходного пространства признаков

при диагностике МЗ для создания информационной структуры БД КСППР в медицине.

Формализация исходных данных. Исходные данные при диагностике МЗ предоставлены Харьковским специализированным медико-генетическим центром (ХСМГЦ). Для анализа были отобраны 145 больных с подозрением на МЗ. В ходе комплексного обследования в ХСМГЦ у них установлено наличие разных форм нарушения биоэнергетического обмена (МЗ), которые включали органические ацидурии, нарушение окисления жирных кислот, нейро-желудочно-кишечную энцефалопатию (MNGIE), синдром MELAS, синдром MERRF, синдром Кернса-Сейра, нейропатию Лебера, болезнь Альцгеймера.

Особенностью исходных данных при этих заболеваниях является наличие большого объема информации, представленной в слабоструктурированном или неструктурированном виде. При этом многие признаки носят описательный характер. Таким образом, анализ исходного пространства признаков показал, что без предварительной формализации полученной информации невозможно создание специализированной БД.

Пусть каждый пациент представляет собой объект ω_i ($\overline{i=1,N}$, N – количество больных) в многомерном пространстве признаков. Пространство признаков порождается множеством признаков X, из элементов которого формируются вектора признаков. В результате каждый объект ω_i в пространстве признаков описывается вектором $\vec{x}_i^{\omega} = (x_{i1}, x_{i2}, \ldots, x_{im})$, а из совокупности объектов ω_i ($i = \overline{1,N}$) формируется таблица экспериментальных данных (ТЭД) типа "объект – признак" (табл. 1).

Таблица 1

Таблица экспериментальных данных

Объекты (пациенты)		Исходные признаки					
		x_1	x_2	$\cdots$	x_j	$\cdots$	x_m
ω_1	$\vec{x}_1^{\omega}$	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_i	$\vec{x}_i^{\omega}$	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{im}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_N	$\vec{x}_N^{\omega}$	x_{N1}	x_{N2}	$\cdots$	x_{Nj}	$\cdots$	x_{Nm}

Для компьютерной обработки экспериментальных данных необходимо, чтобы признаки x_j ($j = \overline{1,m}$) были выражены в числовой, ординальной или номинальной шкалах. Поэтому на первом этапе формализации исходных данных все признаки, имеющие описательный характер, предлагается

разбивать на признаки, которые можно измерить в дихотомической шкале, являющейся частным случаем номинальной шкалы.

На втором этапе формализации исходных данных предлагается выполнить разбивку исходного множества признаков X на непересекающиеся подмножества X_k таким образом, что $\bigcup_{k=1}^{K} X_k = X$, $X_k \cap X_l = \varnothing$, $k,l = \overline{1,K}$, $k \neq l$. В результате были выделены следующие подмножества признаков: X_1 – фенотип, X_2 – диагнозы, X_3 – лабораторные исследования, X_4 – молекулярные исследования. На следующем этапе формализации каждое из подмножеств X_k, в свою очередь, разбивается на подмножества признаков X_k^p. Такая разбивка производится до тех пор, пока подмножества признаков X_k^p не будут содержать однородные по смыслу признаки, исходя из логики дальнейшей обработки экспериментальных данных.

Таким образом, для формализации исходных данных была предложена следующая разбивка исходного пространства признаков на подмножества. Подмножество признаков по фенотипу X_1 было разбито на следующие подмножества: X_1^1 – характеристики состояния кожи (28 признаков); X_1^2 – характеристики состояния ногтей (7 признаков); X_1^3 – характеристики состояния волос (6 признаков); X_1^4 – характеристики состояния подкожной клетчатки (3 признака); X_1^5 – характеристики состояния мышц (4 признака); X_1^6 – характеристики внешнего вида черепа (22 признака); X_1^7 – характеристики внешнего вида лица (8 признаков); X_1^8 – характеристики внешнего вида ушных раковин (12 признаков); X_1^9 – характеристики области глаз и глазного яблока (33 признака); X_1^{10} – характеристики внешнего вида носа (15 признаков); X_1^{11} – характеристики губ и полости рта (15 признаков); X_1^{12} – характеристики верхней и нижней челюстей (8 признаков); X_1^{13} – характеристики зубов (5 признаков); X_1^{14} – характеристики языка (6 признаков); X_1^{15} – характеристики неба (6 признаков); X_1^{16} – характеристики внешнего вида шеи (5 признаков); X_1^{17} – характеристики внешнего вида грудной клетки (9 признаков); X_1^{18} – характеристики состояния позвоночника (7 признаков); X_1^{19} – характеристики внешнего вида живота, таза и ягодиц (10 признаков); X_1^{20} – характеристики внешнего вида верхних конечностей (26 признаков); X_1^{21} – характеристики внешнего вида нижних конечностей (31 признак). Подмножество признаков по диагнозам X_2 было разбито на следующие подмножества: X_2^1 – дыхательная система; X_2^2 – зрительная система; X_2^3 – мочеполовая система; X_2^4 – нервная система; X_2^5 – опорно-двигательная система; X_2^6 – пищеварительная система; X_2^7 – покровная система; X_2^8 – репродуктивная система; X_2^9 – сердечно-сосудистая система; X_2^{10} – слуховая система; X_2^{11} – эндокринная система. Подмножество признаков по лабораторным исследованиям X_3 было разбито на следующие

подмножества: X_3^1 – скрининг-тест мочи; X_3^2 – биохимический анализ крови; X_3^3 – биохимический анализ мочи; X_3^4 – тонкослойная хроматография (ТСХ) аминокислот (АК) крови; X_3^5 – ТСХ АК мочи; X_3^6 – ТСХ углеводов мочи. Разбивка множества X_4 не выполнялась, т.к. оно содержит однородные признаки, характеризующие состояние полиморфизмов генов 677 C→T MTHFR и 66 A→G MTRR.

Признаки, входящие в подмножества X_1^i ($\overline{i=1,21}$), X_2^j ($\overline{j=1,11}$) и X_3^1, измеряются в дихотомической шкале, поэтому принимаем значение 0 – отсутствие признака, значение 1 – наличие. Значения показателей подмножеств X_3^k ($\overline{k=2,6}$), полученные в результате лабораторных исследований, измеряются в количественной шкале, однако нормы по этим показателям зависят от возраста пациента. Поэтому для удобства совместного анализа данных больных, принадлежащих различным возрастным группам, предлагается привести признаки подмножеств X_3^k ($\overline{k=2,6}$) с учетом возрастных норм к ординальной шкале. При этом значение 0 принимает признак, если показатель в норме, положительное значение – если показатель превышает норму, отрицательное – если показатель ниже нормы. Величина признака показывает степень отклонения показателя от нормы.

Таким образом, предложенное представление исходного пространства признаков в виде иерархической структуры непересекающихся подмножеств позволило формализовать исходные данные при диагностике МЗ.

Разработка информационной структуры БД. Схема данных специализированной БД КСППР для диагностики МЗ с учетом разработанной иерархической структуры непересекающихся подмножеств признаков представлена на рис. 2. Представление исходного пространства признаков в виде иерархической структуры непересекающихся подмножеств позволило выделить ряд справочников, входящих в информационную структуру БД проектируемой КСППР. Каждый справочник представляет собой таблицу, содержащую признаки описанных выше подмножеств. Для реализации разбивки множеств X_1 и X_2 на подмножества были организованы дополнительные справочники, представляющие собой таблицы, в которых хранятся названия подмножеств X_k^p.

Рассмотрим организацию справочников на примере хранения признаков множества X_2. В таблице bolezni (справочник множества X_2) имеются следующие поля (см. рис. 2): id_bolezni – уникальный ключ; name – название диагноза; id_sys – ключ для связи со справочником по системам организма; onkonkolog – признак принадлежности диагноза к онкологическому заболеванию. В таблице bolezni_group, являющейся справочником названий подмножеств X_2^p (см. рис. 2), хранится уникальный ключ (поле id_system) и

перечень систем организма (поле system). Для добавления новой болезни в справочник необходимо ввести ее название, выбрать систему организма, к которой она относится, и указать, является ли этот диагноз онкологическим.

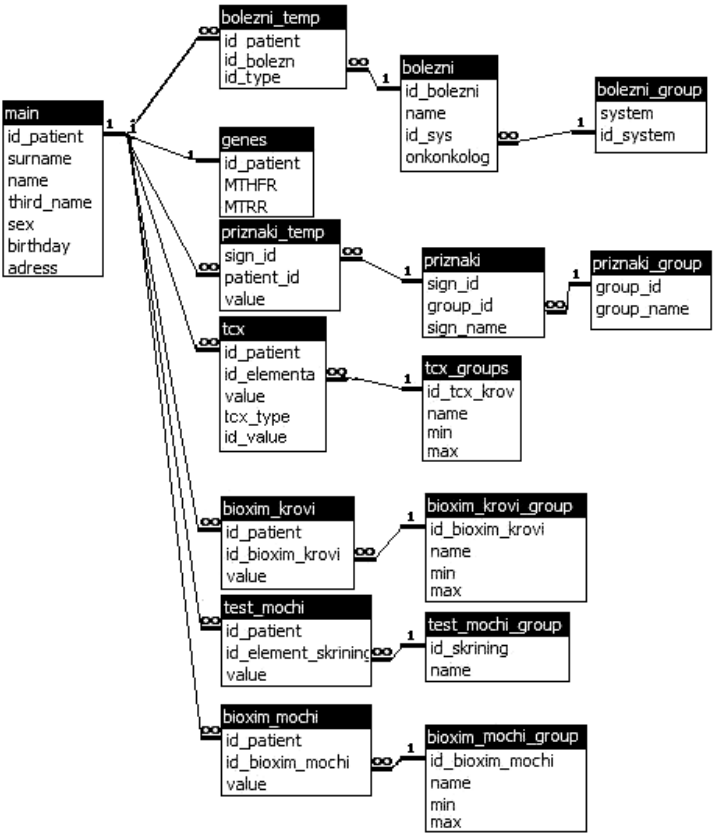

Рис. 2. Схема данных специализированной БД

Надо отметить, что особенностью подмножеств признаков X_1 и X_2 является то, что с увеличением объема выборки (то есть с добавлением новых пациентов) количество признаков подмножеств X_1 и X_2 может увеличиться. Например, у нового пациента может появиться новый признак в фенотипе или новый диагноз в родословной. Организация БД, представленная на рис. 2, позволяет без изменения схемы данных не только увеличивать объем выборки, но и увеличивать количество признаков за счет добавления новых записей в соответствующие справочники.

Предварительный анализ ТЭД. Согласно общей схеме анализа экспериментальных данных (см. рис. 1) после этапов сбора и формализации необходим этап предварительного анализа внутренней структуры ТЭД. Как было отмечено выше, все признаки x_j, описывающие объекты ω_i, являются либо ординальными, либо дихотомическими. При этом для ординальных признаков шкалы измерения могут быть симметричными относительно нуля, если параметр может быть как выше, так и ниже нормы, и несимметричными, если параметр может быть только выше нормы. Поэтому необходимо преобразовать значения признаков таким образом, чтобы они все были измерены в однотипных шкалах. В качестве такого преобразования предлагается выполнить сдвиг и нормировку признаков. В результате преобразования все признаки будут измеряться в диапазоне $x_{ij}^* \in [0,1]$:

$$x_{ij}^* = \frac{x_{ij} - \min x_j}{\max x_j - \min x_j},$$

где x_{ij}, x_{ij}^* – исходное и преобразованное значение признака x_j, измеренного у объекта ω_i; $\max x_j$, $\min x_j$ – максимальное и минимальное значение признака x_j.

Структура экспериментальных данных отражается посредством двух основных категорий взаимоотношений между элементами ТЭД – категорий сходства и различия. Сходство и различие объектов ТЭД отражается с помощью матрицы удаленности объектов $\mathbf{D} = \{d_{il}\}_{i,l=1}^{N}$ [10]. В качестве меры различия объектов ТЭД предлагается использовать расстояние Хемминга:

$$d_{il} = \sum_{j=1}^{m} \left| x_{ij}^* - x_{lj}^* \right|,$$

где x_{ij}^*, x_{lj}^* – преобразованные значения признака x_j, измеренные у объектов ω_i и ω_l соответственно.

Для оценки существенности связи двух номинальных признаков на основе анализа таблиц сопряженности (табл. 2) используются методы сравнения эмпирических и теоретических частот по Брандту и Снедекору. В табл. 2 приняты следующие обозначения: n_{fg} – число пациентов, у которых признак x_{ik}^* относится к классу f и одновременно признак x_{ij}^* относится к классу g; $n_{f\bullet}$ – общее число пациентов, у которых признак x_{ik}^* относится к классу f; $n_{\bullet g}$ – число пациентов, у которых признак x_{ij}^* относится к классу g; l, p – число градаций признаков x_k и x_j соответственно; N – длина выборки.

Таблица 2

Таблица сопряженности номинальных признаков общего вида

Градации (классы) признака		$x^*_{ij}(i=\overline{1,N})$					
		1	…	g	…	p	
$x^*_{ik}(i=\overline{1,N})$	1	n_{11}	…	…	…	n_{1p}	$n_{1\bullet}$
	…	…	…	…	…	…	…
	f	…	…	n_{fg}	…	…	$n_{f\bullet}$
	…	…	…	…	…	…	...
	l	n_{l1}	…	…	…	n_{lp}	$n_{l\bullet}$
		$n_{\bullet 1}$	…	$n_{\bullet g}$	…	$n_{\bullet p}$	N

Вычисление коэффициента квадратичной сопряженности основывается на расчете критерия $\chi^2_{кр}$, оценивающего меру близости по всем ячейкам таблицы сопряженности [7]:

$$\chi^2_{кр} = \sum_{f=1}^{l}\sum_{g=1}^{p} \frac{\left(n_{fg} - \dfrac{n_{f\bullet}n_{\bullet g}}{N}\right)^2}{\dfrac{n_{f\bullet}n_{\bullet g}}{N}}.$$

Сходство и различие признаков ТЭД отражается с помощью матрицы связей признаков $\mathbf{S} = \{s_{jk}\}_{j,k=1}^{m}$ [10]. В качестве меры связи признаков ТЭД предлагается использовать коэффициент квадратичной сопряженности:

$$s_{jk} = \begin{cases} \dfrac{\chi^2_{кр}}{2N}, & \text{если } \chi^2_{кр} < \chi^2_{1-\alpha}(v), \\ 0, & \text{в противном случае,} \end{cases}$$

где $\chi^2_{1-\alpha}(v)$ – табличное значение распределения хи-квадрат с числом степеней свободы $v = (l-1)(p-1)$.

С помощью полученных таблиц $\mathbf{D} = \{d_{il}\}_{i,l=1}^{N}$ и $\mathbf{S} = \{s_{jk}\}_{j,k=1}^{m}$ выполняется анализ внутренней структуры ТЭД, который показывает наличие кластеров объектов в заданном пространстве признаков и наличие связанных признаков.

Выводы. В данной работе выполнены этапы сбора, формализации и

предварительного анализа исходных признаков и предложен способ создания информационной структуры БД КСППР для диагностики МЗ, который позволяет выполнять добавление новых признаков и обновление уже существующих признаков без изменения структуры БД.

Список литературы: **1**. *Гречаніна Ю.Б.* Клінічно-генетична і молекулярна діагностика мітохондріопатій // Ультразвукова перинатальна діагностика. – 2005. – № 18. – С. 148–163. **2.** *Wallace C.D., Brown D.M., Lott T.M.* Mitochondrial Genetics // Gene. – 1999. – P. 277– 317. **3.** *Гречанина Е.Я.* Проблемы клинической генетики. – Харьков: КВАДРАТ, 2003. – 420 с. **4.** *Гречанина Ю.Б., Васильева О.В.* Клинические "маски" митохондропатий // Медицина третього тисячоліття: збірник тез. – Харків, 2007. – С. 80. **5.** *Гречанина Е.Я.* Молекулярная медицина: реальность и перспективы. – Харьков, 2007. – 120 с. **6.** *Гречанина Ю.Б.* Стандарти для визначення мітохондропатій // Ультразвукова перинатальна діагностика. – 2003. – № 16. – С. 131–145. **7.** *Гланц С.* Медико-биологическая статистика. – М.: Практика, 1998. – 459 с. **8.** *Александров В.В., Алексеев А.И., Горский Н.Д.* Анализ данных на ЭВМ (на примере системы СИТО). – М.: Финансы и статистика, 1990. – 192 с. **9.** *Лбов Г.С.* Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука, 1981. – 157 с. **10.** *Дюк В.А.* Компьютерная психодиагностика. – СПб.: Братство, 1994. – 364 с.

УДК 61:004.8

Створення інформаційної структури бази даних комп'ютерної системи підтримки прийняття рішень для діагностики мітохондріальних захворювань / Бойко Д.О., Васильєва О.В., Галкін Д.О., Гречаніна Ю.Б., Поворознюк А.І., Філатова Г.Є. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2009. – № 13. – С. 14 – 22

У роботі розглядається задача формалізації вхідних даних при діагностиці мітохондріальних захворювань. В якості вхідних ознак було розглянуто фенотип, біохімічні дослідження крові і сечі, а також захворювання в родоводі і супутні діагнози. В результаті розроблені довідники для ведення БД. Робота ведеться спільно з Харківським спеціалізованим медико-генетичним центром. Іл.: 2. Табл.: 2. Бібліогр.: 10 назв.

Ключові слова: формалізація вхідних даних, мітохондріальні захворювання, фенотип, біохімічні дослідження крові і сечі, база даних.

UDC 61:004.8

Creation of database informative structure of decisions acceptance support computer system for mitochondrial diseases diagnostics / Boyko D.A., Vasylieva O.V., Galkin D.A., Grechanina J.B., Povoroznyuk A.I., Filatova A.E. // Herald of the National Techical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2009. – №. 13. – P. 14 – 22.

The task of source data formalization in mitochondrial diseases diagnosing is considered in work. As source signs were considered phenotype, biochemical examination of blood and urine, and genealogical diseases and attendant diagnoses. As a result reference books in operating databases were worked out. The work is conducted with Kharkov specialized medico-genetic centre. Figs: 2. Tabl: 2. Refs: 10 titles.

Key words: source data formalization**,** mitochondrial diseases, phenotype, biochemical examination of blood and urine, databases.

Поступила в редакцию 19.05.2009