Институт системного анализа
Федеральный исследовательский центр
“Информатика и управление”
Российской академии наук

Автоматизация извлечения информации из научных статей в области биомедицины

Роман Суворов
м.н.с. лаб 0-7
ИСА ФИЦ ИУ РАН

+7 (499) 135 04 63

rsuvorov@isa.ru

117312, Москва,
пр. 60-летия Октября 9

Москва, 2017

Актуальность

- Составление обзоров – необходимый этап в исследовательском процессе
- > 1,5 млн. статей в год

- Нужны инструменты

Систематический обзор

1)Формирование протокола

- Основной вопрос
- Критерии отбора статей
- Структура извлекаемой информации

2)Сбор первичных материалов

- Клинические исследования

3)Извлечение структурированной информации

- Вручную или автоматизированно

4)Критический анализ и обобщение

Систематический обзор

1)Формирование протокола

- Основной вопрос
- Критерии отбора статей
- Структура извлекаемой информации

2)Сбор первичных материалов

- Клинические исследования

3)Извлечение структурированной информации

- Вручную или **автоматизированно**

4)Критический анализ и обобщение

Цели доклада

- Извлечение информации из текстов на естественном языке
- Основные задачи и сложности
- Некоторые авторские экспериментальные результаты
- Работа в процессе

Объекты

- 0) Paper confirmation
  - 0) Paper confirmation #23372
    - Cancer Stem Cell: Yes
      - cancer stem cells
      - CSCs
      - CSCs
    - Cell sorting: Yes
      - κ ow cytometer
      - cyto κ uorometric cell sepa
      - sorted with a FACS Aria
      - cyto κ uorometric sorting
      - flow cytometry
    - Marker Name: Yes
      - CD44
      - CD54
      - CD45
      - CD31
    - Human: Yes
      - patients
      - human
      - patient
      - patients
      - patients
      - patients
    - Decision: Yes

Содержимое

Cell Research (2012) 22:248-258.

www.nature.com/cr

ORIGINAL ARTICLE

Identification and expansion of cancer stem cells in tumor tissues and peripheral blood derived from gastric adenocarcinoma patients

Tie Chen[1, *], Kun Yang[2, *], Jianhua Yu[1], Wentong Meng[1], Dandan Yuan[3], Feng Bi[3], Fang Liu[1], Jie Liu[2], Bing Dai[2], Xinzu Chen[2], Fang Wang[2], Fan Zeng[1], Hong Xu[1], Jiankun Hu[2], Xianming Mo[1]

[1]Laboratory of Stem Cell Biology, [2]Department of Gastrointestinal Surgery and [3]Oncology, State Key Laboratory of Biotherapy, West China Hospital, West China Medical School, Sichuan University, Chengdu, Sichuan 610041, China

Gastric cancer is the fourth most common cancer worldwide, with a high rate of death and low 5-year survival rate. To date, there is a lack of efficient therapeutic protocols for gastric cancer. Recent studies suggest that cancer stem cells (CSCs) are responsible for tumor initiation, invasion, metastasis, and resistance to anticancer therapies. Thus, therapies that target gastric CSCs are attractive. However, CSCs in human gastric adenocarcinoma (GAC) have not been described. Here, we identify CSCs in tumor tissues and peripheral blood from GAC patients. CSCs of human GAC (GCSCs) that are isolated from tumor tissues and peripheral blood of patients carried CD44 and CD54 surface markers, generated tumors that highly resemble the original human tumors when injected into immunodeficient mice, differentiated into gastric epithelial cells *in vitro*, and self-renewed *in vivo* and *in vitro*. Our findings suggest that effective therapeutic protocols must target GCSCs. The capture of GCSCs from the circulation of GAC patients also shows great potential for identification of a critical cell population potentially responsible for tumor metastasis, and provides an effective protocol for early diagnosis and longitudinal monitoring of gastric cancer.

Keywords: cancer stem cells; gastric adenocarcinoma; CD44; CD54; circulating tumor cells
Cell Research (2012) **22**:248-258. doi:10.1038/cr.2011.109; published online 5 July 2011

Связанные работы

- Анализ клинических текстов
  - Задачи:
    - Заболевания, симптомы, временная информация, условия, субъект и т.п.
  - Программные средства:
    - cTAKES, MedIE, MedEx, MetaMap
  - Методы:
    - правила, CRF, сопоставление со словарями

Связанные работы

- Анализ научных статей
  - Задачи:
    - Отношения ген-заболевание, белок-белок, лекарство-реакция и т.п.
  - Программные средства:
    - ExaCT, TEES, ABNER, GenIE
  - Методы:
    - классификация предложений + регулярные выражения, SVM, CRF

Типы источников и подструктур

- Источники
  - Текст
  - Таблицы
- Подструктуры
  - Идентификаторы исследуемых объектов — «Пациент 1» и т.п.
  - Именованные сущности — наименования заболеваний, органов, идентификаторы фенотипов и т.п.
  - Размерные числовые величины — дозировки, концентрации.
  - Отношения «объект-свойство».
- Нормализация выделенных фрагментов
  - Классификация
  - Выделение числовой величины

Проблемы

- Сложность адаптации под новую структуру извлекаемой информации
- Необходимость больших размеченных корпусов
- Неуниверсальность признакового пространства

Цели и задачи исследования

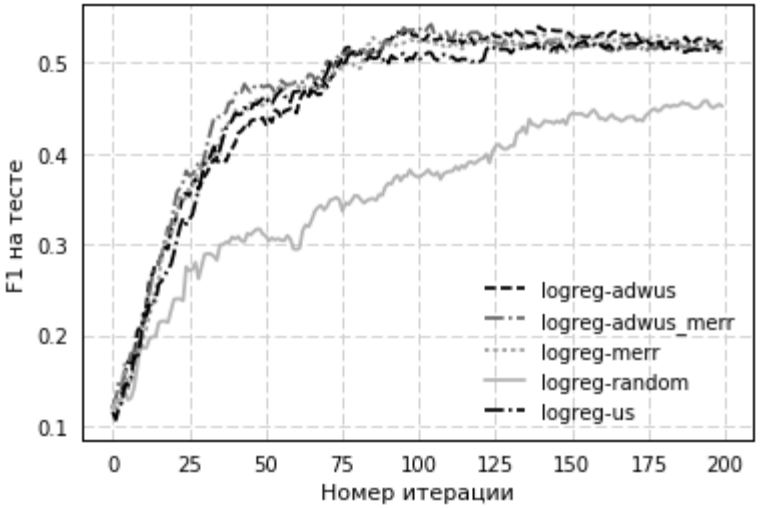
- Снижение трудозатрат
  - на разметку статей (эксперты предметной области)
  - на настройку алгоритмов (инженер по данным)
- Эффективное использование существующих инструментов
- Максимальное использование имеющейся в статье информации

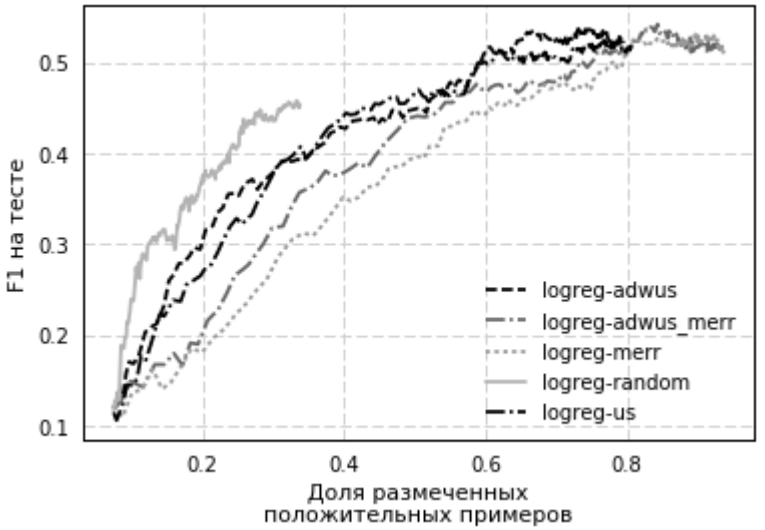
- Реализация
  - Активное обучение
  - Графовое представление текста и результатов его разбора
  - Извлечение и анализ таблиц

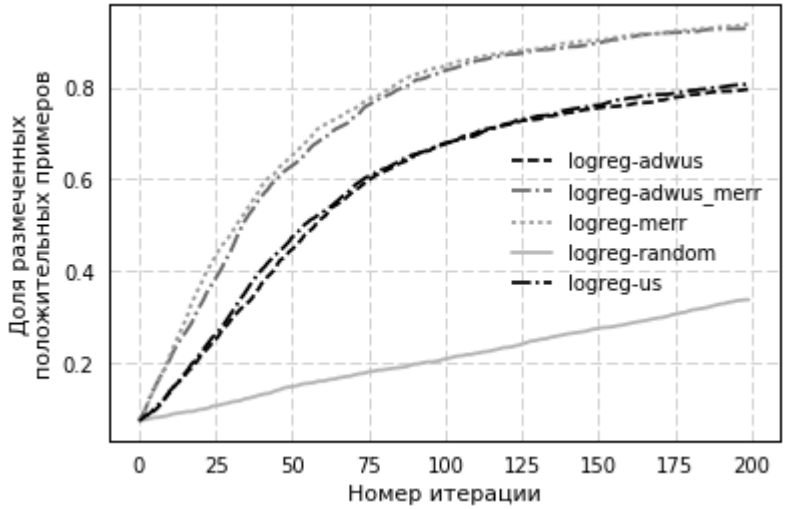
Активное обучение

- Разметка начального набора документов
- Итерационная процедура:
  - Обучение классификаторов на размеченных данных
  - Выбор из ***неразмеченных*** данных наиболее ***интересных*** примеров с точки зрения обученных классификаторов
  - Ручная разметка ***только наиболее интересных*** примеров
- Критерий ***интересности*** определяется стратегией выбора примеров:
  - по неуверенности
  - с учетом плотности
  - по степени несогласия ансамбля

Активное обучение







- Экономия трудозатрат
  - на >80%
  - ~ в 6 раз

Графовое представление текстов

- Цель:
  - Эффективное использование существующих инструментов анализа текстов
  - Упрощение настройки под новую структуру извлекаемой информации
- Граф:
  - Вершины: словоупотребления, элементы связанных онтологий, результаты анализа сторонними инструментами и т.п.
  - Рёбра: следование, синтаксическое подчинение, сем. связи и роли, привязка к онтологиям, отношения, выделенные сторонними инструментами и т.п.
- Построение признакового пространства:
  - Заданные экспертом правила обхода
  - Случайные обходы
  - Сжатые (распределённые) векторные представления

Анализ таблиц

- Подзадачи:
  - Определение области страницы
  - Разбиение на столбцы и строки
- Инструменты: Tabula, pdffigures2
- Ошибки:
  - Включение в область страницы лишних элементов (особенно при двухколоночном форматировании)
- Решение:
  - Разметка обучающей выборки с помощью эвристических правил
  - Обучение классификатора, опирающегося на информацию о содержимом соседних строк, и предсказывающего начало и окончание таблицы

Примеры

- Успешность применения дендритоклеточных вакцин
  - Информация о пациентах
  - Условия, при которых лечение проходит более успешно
- Поверхностные маркеры РСК
  - Информация о поверхностной экспрессии белков клетками
  - Отличительные характеристики РСК

Типы объектов **Удаление свойств разрешено**: Нет +

1) Группа пациентов Добавить свойство ▾ | Создать дубликат | 🗑 **Тип является самостоятельным**: Да

Название	Тип	Описание	Другие свойства	Порядок	Удалить
Количество пациентов	Целое число	Описание	**Значение по умолчанию**: 1	10	🗑
Возраст	Список объектов	Описание	**Встроенный**: Да **Тип значения**: Возраст ➦	20	🗑
Пол	Список объектов	Описание	**Встроенный**: Да **Тип значения**: Пол ➦	30	🗑
Раса	Список объектов	Описание	**Встроенный**: Да **Тип значения**: Раса ➦	40	🗑
Гаплотип	Список объектов	Описание	**Встроенный**: Да **Тип значения**: Гаплотип ➦	50	🗑
Диагноз	Список объектов	Описание	**Встроенный**: Да **Тип значения**: Диагноз ➦	60	🗑
Количество пациентов	Целое число	Количество пациентов, у которых наблюдается именно это значение признака, если представлено в тексте	**Значение по умолчанию**: 0	40	🗑

Адъювант в составе вакцины Добавить свойство ▾ | Создать дубликат | 🗑 **Тип является самостоятельным**: Нет

Название	Тип	Описание	Другие свойства	Порядок	Удалить
Доля пациентов	Вещественное число	Процент пациентов (от 0 до 100), у которых наблюдается именно это значение признака, если представлено в тексте	**Значение по умолчанию**: 100,0	0	🗑
Значение	Номинальное значение	Значение признака	**Значение по умолчанию**: -	0	🗑
Количество пациентов	Целое число	Количество пациентов, у которых наблюдается именно это значение признака, если представлено в тексте	**Значение по умолчанию**: 0	0	🗑

Объекты

- 1) Группа пациентов
  - ALL #2996
    - Количество пациентов: 14
      - Fourteen patients
    - Возраст (Список объектов)
    - Пол (Список объектов)
      - Пол: Пол #12321
        - Доля пациентов: 100.
        - Значение: Male
          - Male
        - Количество пациенто
          - 9
      - Пол: Пол #12322
        - Доля пациентов: 100.
        - Значение: -

Содержимое

AB serum. After 1 hr at 37°C, the blocking medium was discarded and 1 × 10^5 $CD8^+$ T cells/well, purified from patient PBMCs by positive selection with CD8 MicroBeads (Miltenyi, Bergisch Gladbach, Germany), were added. T2 cells (7.5 × 10^4/well), pulsed with the relevant peptides in the presence of β2microglobulin, were added. After a culture period of 20 hr at 37°C, cells were removed and biotinylated anti-IFN-γ MAb (clone 7-B6-1, Mabtech) was added for 2 hr at 37°C. A 100 μl volume of ABC (ABC Vectastain-Elite kit; Vector, Burlingame, CA) was added at a dilution of 1/100 and incubated for 1 hr at room temperature. After unbound complex was removed, peroxidase staining was performed using the substrate 3-amino-9-ethylcarbazole (Sigma). Spots appeared within 4 to 5 min. The color reaction was stopped, and numbers and areas of resulting spots were determined with the use of computer-assisted video image analysis (Herr *et al.*, 1997).

Melanoma-inhibiting activity (MIA) assay

Bosserhoff *et al.* (1997) have demonstrated that MIA can be a useful marker of tumor progression during follow-up of melanoma patients and in monitoring therapy of advanced disease. Human MIA was measured by a 1-step ELISA (Roche, Mannheim, Germany), following the instructions of the manufacturer. The assay is sensitive up to 0.1 ng MIA/ml.

Age (years)	
Range	30–70
Median	50
Performance status (Karnovksy)	
100	7
90	5
80	2
Sex	
Male	9
Female	5
Prior therapy	
Surgery	14
Chemotherapy[1]	3
Chemo-/immunotherapy[2]	9
Immunotherapy[3]	5
Metastatic site	
Skin	2
Lymph node	6
Soft tissue	2
Lung	7
Liver	5
Bone	1

[1]Dacarbazine.–[2]Dacarbazine + vinblastine + cisplatin + IL-2 + IFN-α2a.–[3]IFN-α2a.

Объекты

- P#1 #3016
  - Количество пациентов: 1
    - Patient number
  - Возраст (Список объектов)
  - Пол (Список объектов)
  - Раса (Список объектов)
  - Гаплотип (Список объектов)
    - Гаплотип: Гаплотип #123
      - Доля пациентов: 100.
      - Значение: HLA-A1
        - HLA - A1
      - Количество пациенто
  - Диагноз (Список объектов)
  - Стадия заболевания (Список
  - Индекс ECOG (Список объект

Содержимое

TABLE II – CLINICAL RESULTS

Patient number	HLA-A	Metastatic site	Number of vaccinations	Number of DCs injected	Clinical course (months)
1	HLA-A1	LN, skin	8	5 × 10^6, 1 × 10^7	NC[1] (3)
2	HLA-A2	LN	12	5 × 10^6, 1 × 10^7	NED[2] (19)
3	HLA-A1,A2	Lung	4	5 × 10^6	PD
4	HLA-A2	LN, soft tissue	7	5 × 10^6, 1 × 10^7	NC (4)
5	HLA-A2	Lung, liver	8	1 × 10^7	NC (8)
6	HLA-A1	Lung, liver	4	1 × 10^7	PD
7	HLA-A2	Soft tissue	4	1 × 10^7	PD
8	HLA-A1	Skin	8	1 × 10^7, 5 × 10^7	MR[3] (6)
9	HLA-A2	Lung, liver, bone	4	1 × 10^7	NC (3)
10	HLA-A2	Lung	8	1 × 10^7	NC (6)
11	HLA-A1	Lung, liver, LN	4	1 × 10^7	PD
12	HLA-A2	Lung, LN	4	5 × 10^7	PD
13	HLA-A1	Liver, LN	4	5 × 10^7	PD
14	HLA-A2	Lung	4	5 × 10^7	NC (6^+)

[1]NC, no change.–[2]NED, no evidence of disease.–[3]MR, minor response [complete regression of s.c. metastasis (see Fig. 2), stable disease of another cutaneous metastasis].

Заключение

- Разрабатываются интегрированные программные средства для интерактивного анализа текстов
  - снижение трудоемкости построения системы для извлечения информации
  - использование существующих инструментов
- Ряд сложностей:
  - простой для адаптации алгоритм построения признакового пространства, учитывающий особенности задачи
  - повышение качества извлечения информации

Спасибо за внимание!

Суворов Роман
rsuvorov@isa.ru
м.н.с. лаб. 0-7
ФИЦ ИУ РАН

Анализ извлечённых данных

- Традиционные методики мета-анализов
  - Forest plot
- Методы машинного обучения и анализ построенных моделей
  - Случайные леса решающих деревьев
    - Джини, энтропия
  - Локально-линейная аппроксимация нелинейных разделяющих поверхностей