Image tasks @ CLEF-IP 2012

Mihai Lupu, Florina Piroi, Allan Hanbury, Igor Filippov, Alan Sexton

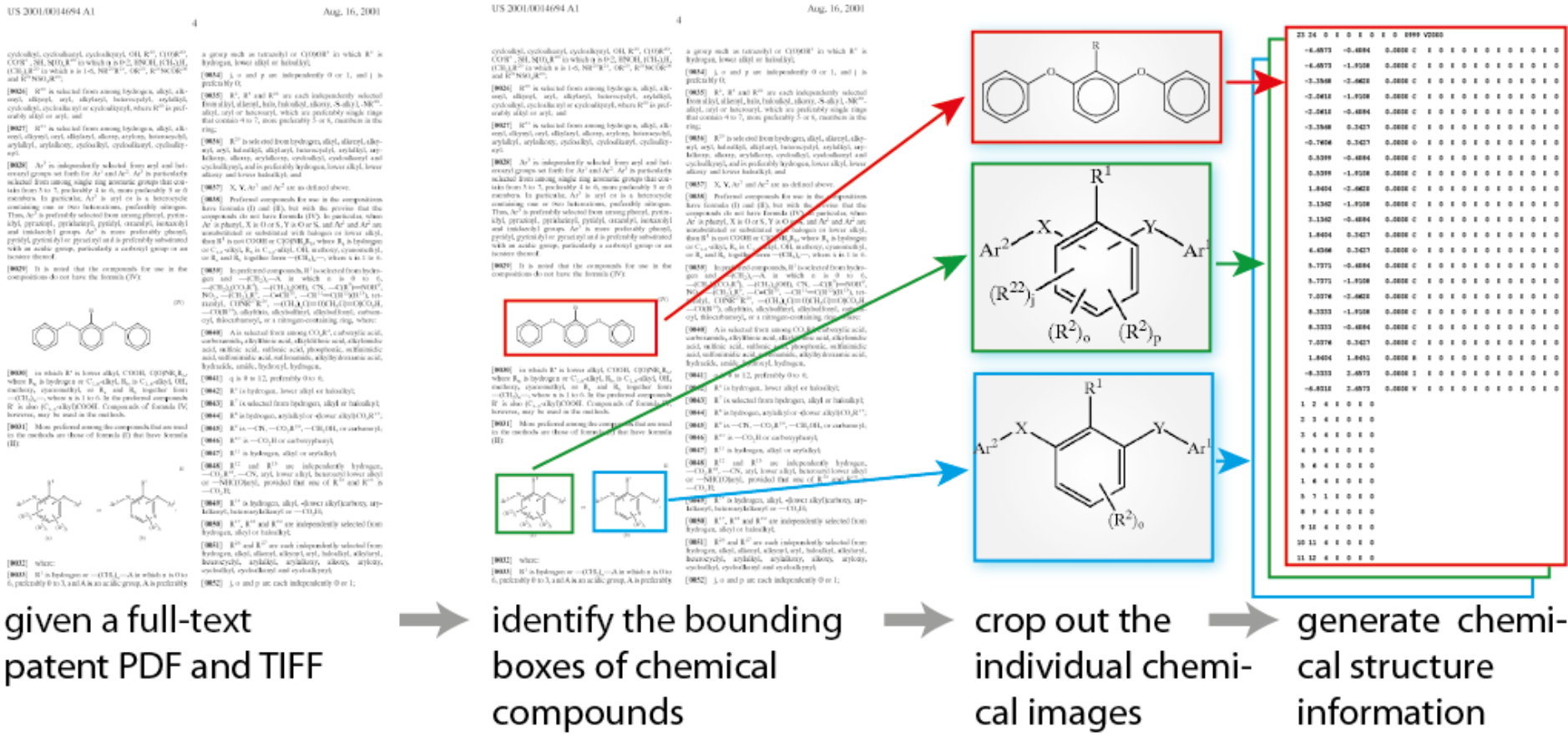
Overview

- Chemical structure recognition results
- Flowchart recognition results
- Next steps

Chemical Structure Evaluation

Mihai Lupu
Igor Filippov
Alan Sexton

Chemical structure recognition



Chemical structure recognition

- Good results for both tasks
  - Segmentation

Tolerance	Precision	Recall	F_1
0	0.70803	0.68622	0.69696
10	0.79311	0.76868	0.78070
20	0.82071	0.79543	0.80787
40	0.86696	0.84025	0.85340
55	0.88694	0.85962	0.87307

  - Recognition

	Automatic Set			Manual Set			Total		
	#Structures	Recalled	%	#Structures	Recalled	%	#Structures	Recalled	%
saic	865	761	88%	95	38	40%	960	799	83%
uob-1	865	832	96%	95	44	46%	960	876	91%
uob-2	865	821	95%	95	56	59%	960	877	91%
uob-3	865	821	95%	95	44	46%	960	865	90%
uob-4	865	832	96%	95	54	57%	960	886	92%

Chemical structure recognition

- Good results for both tasks
  - S
  - F

Caveat

The test structures have been pre-selected to have a INCHI representation
(i.e. no Markush, no 'fancy stuff')

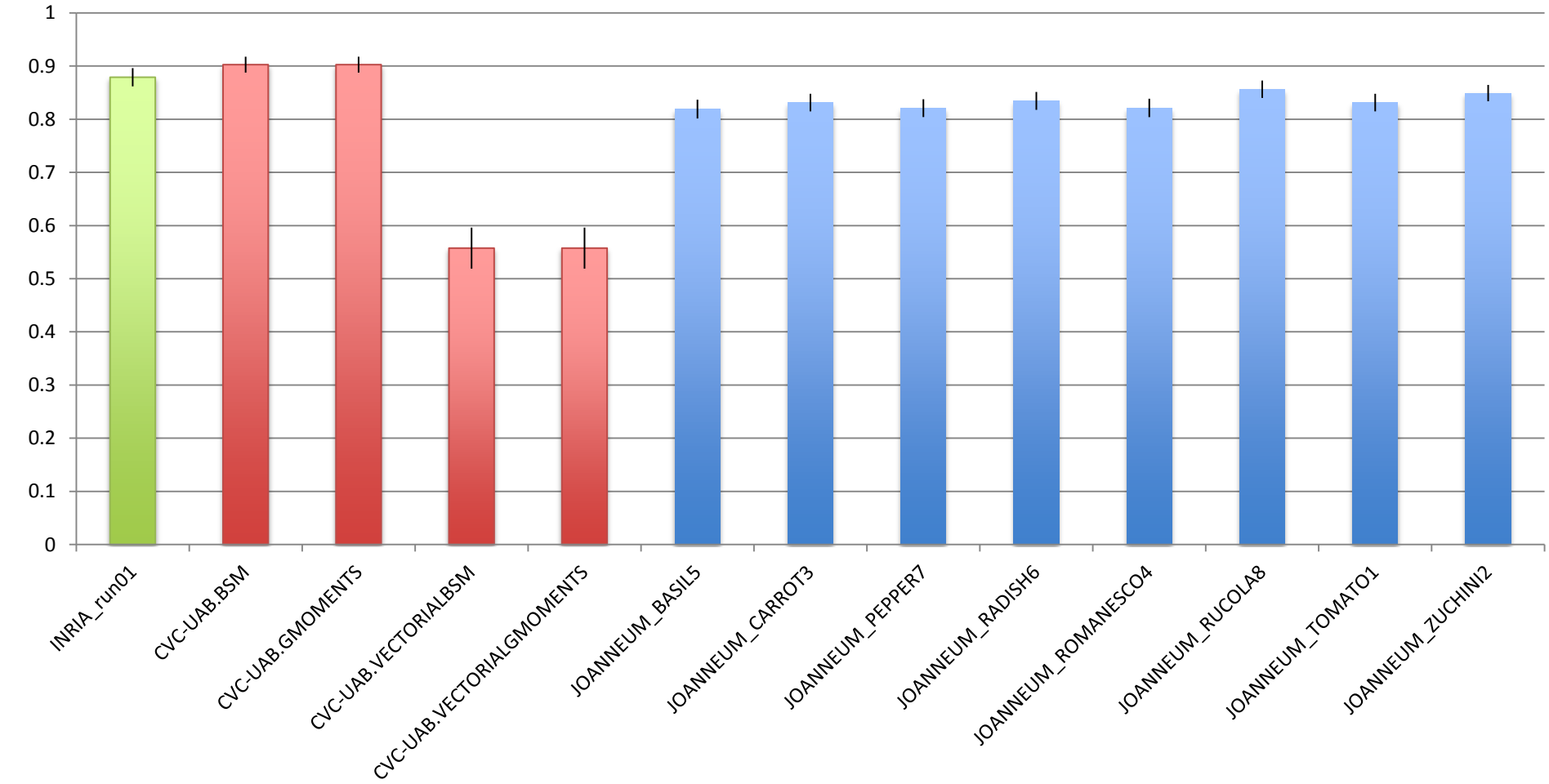
saic									
uob-1	865	832	96%	95	44	46%	960	876	91%
uob-2	865	821	95%	95	56	59%	960	877	91%
uob-3	865	821	95%	95	44	46%	960	865	90%
uob-4	865	832	96%	95	54	57%	960	886	92%

Flowchart Recognition Evaluation

Mihai Lupu
Florina Piroi
Allan Hanbury

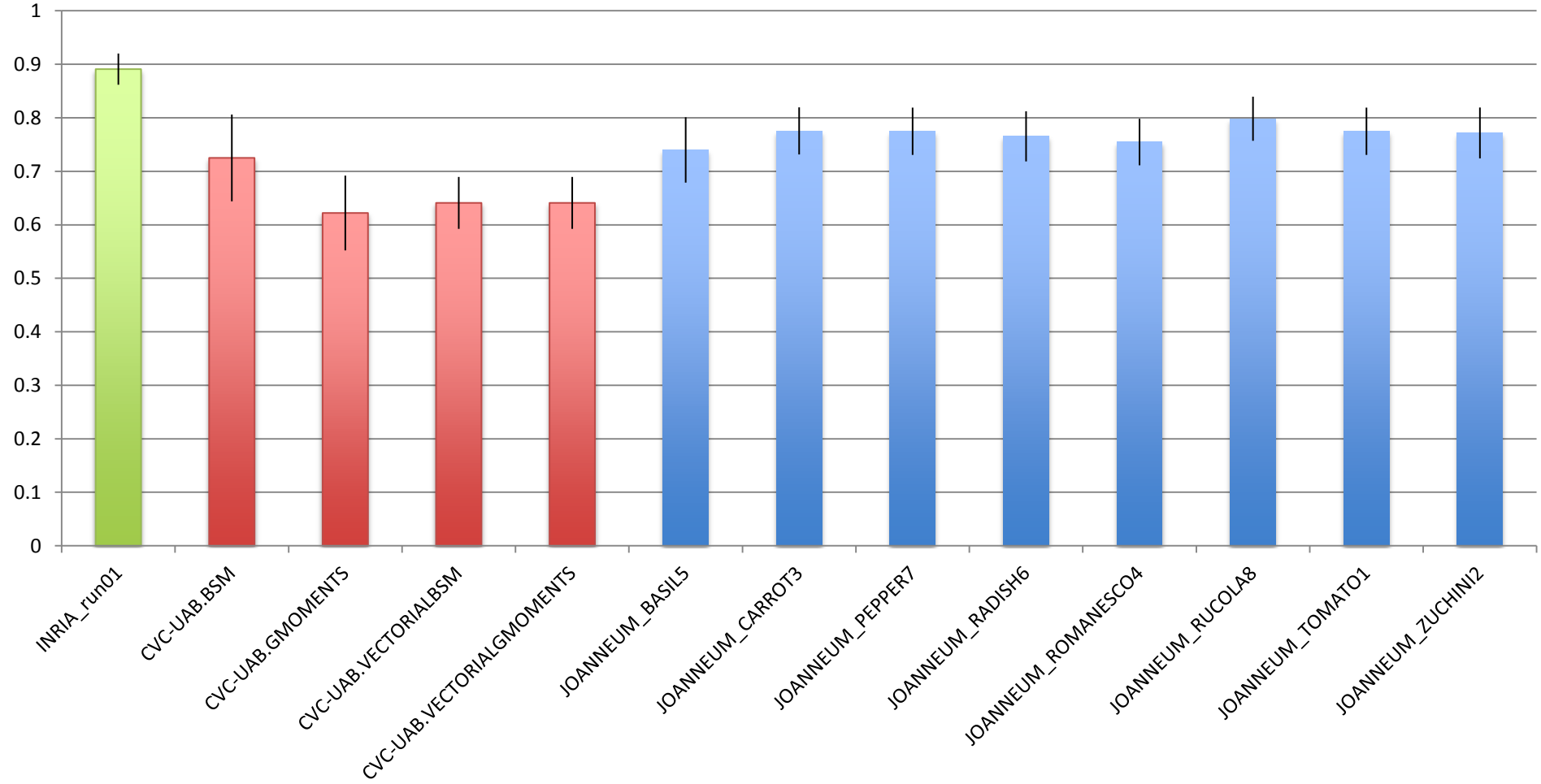
Results (so far)

- Most common sub-graph



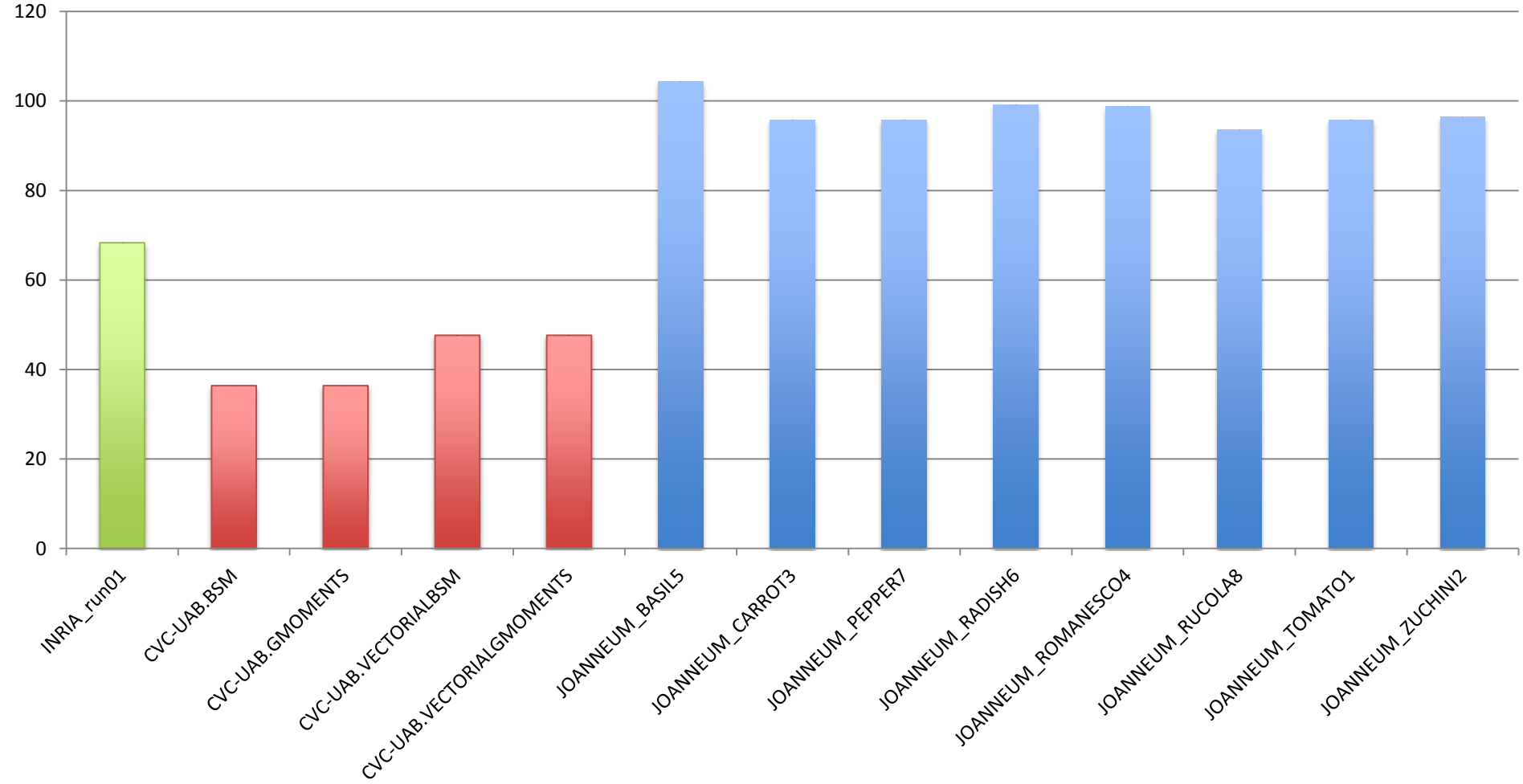
Results (so far)

- Node type match



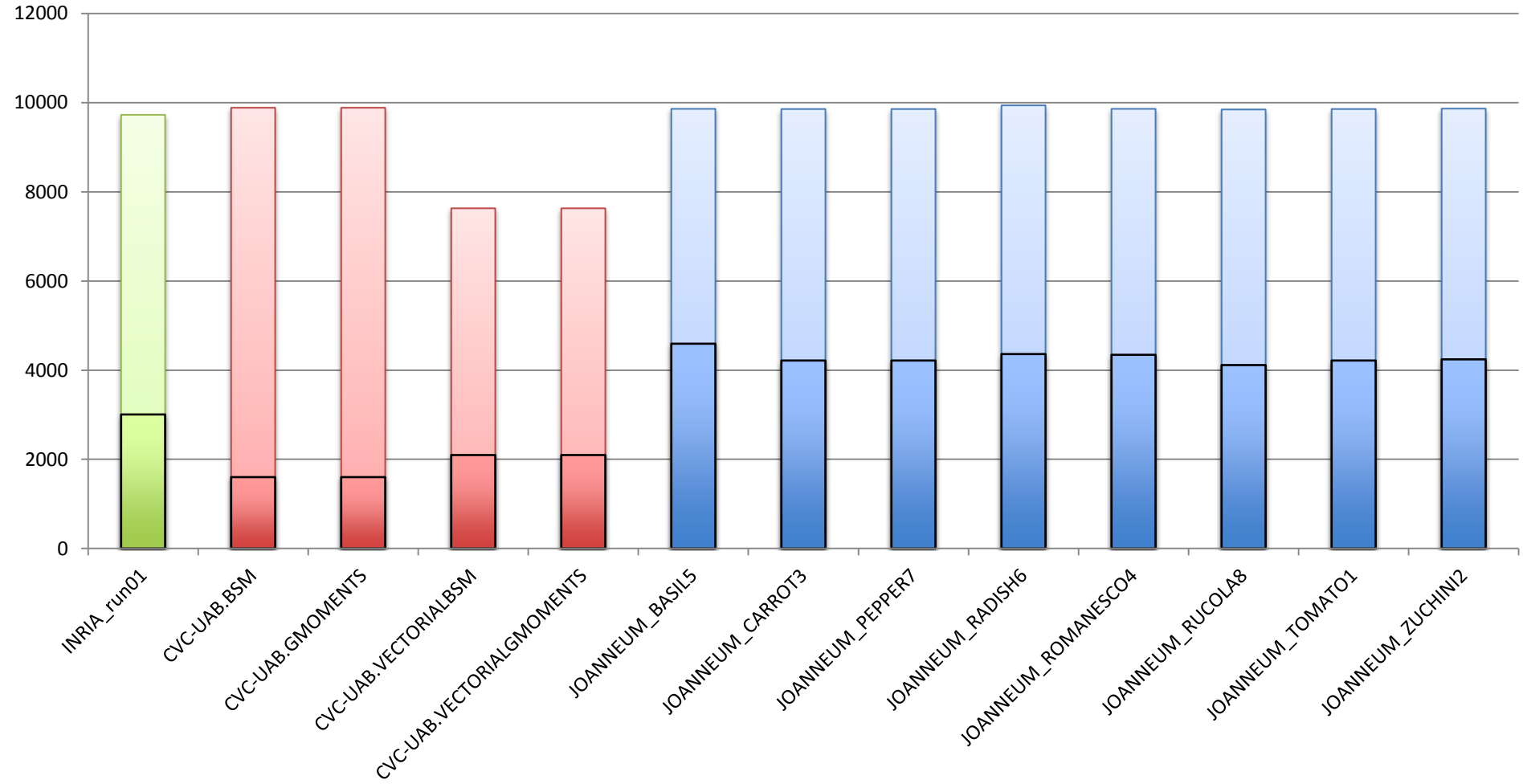
Results (so far)

- Text recognition (Edit Distance) - average



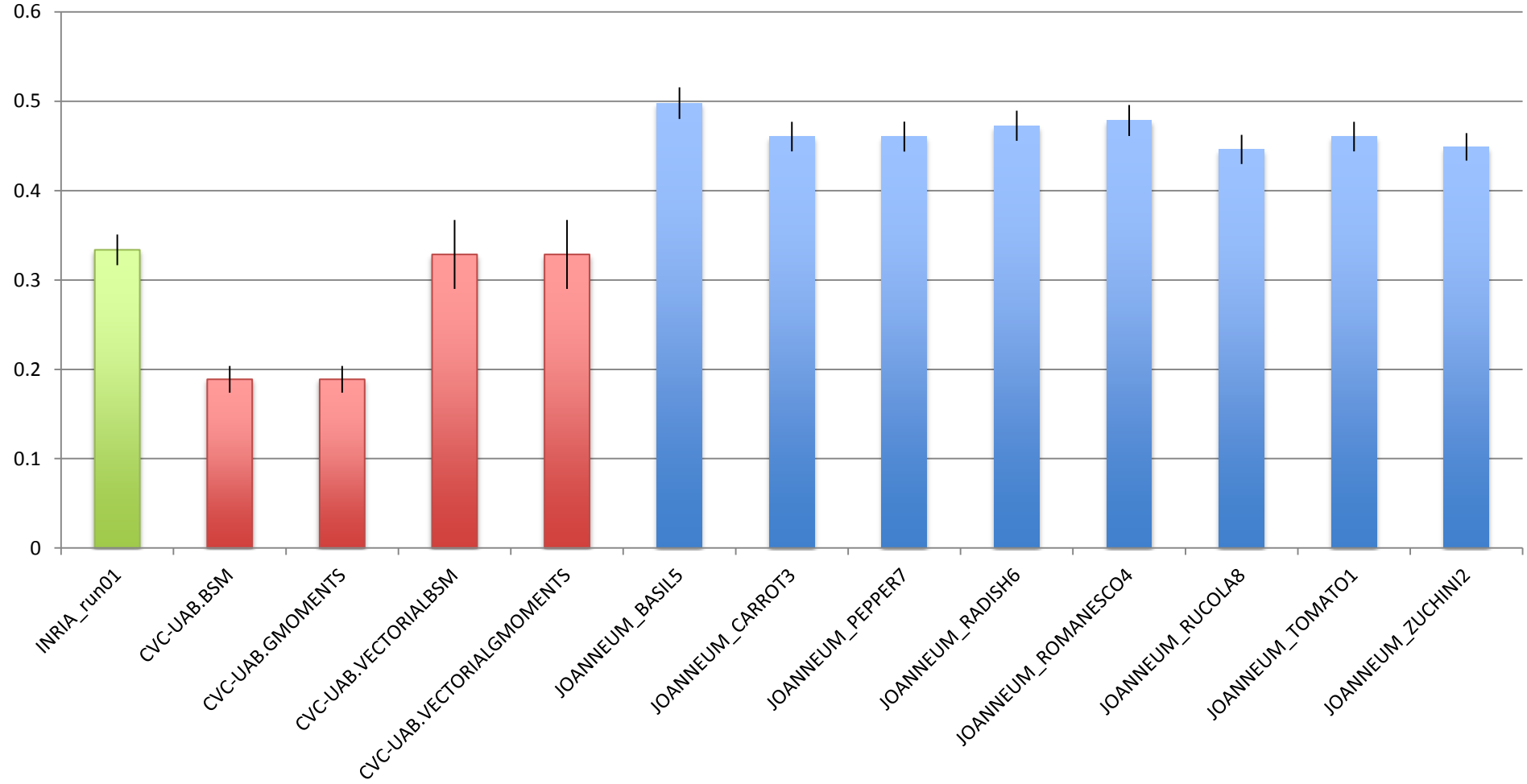
Results (so far)

- Text recognition (Edit Distance) - sum



Results (so far)

- Text recognition (Edit Distance) - normalized



Caveat

- 44 of 100 topics evaluated
  - Only those where a score was computed for all participating runs
    - 'selected' by the algorithm (by not finishing in reasonable time or crashing)

Number of nodes		
	Flowcharts in results	Flowcharts not in results
Min	6	10
Max	21	51
Median	13	22
Average	12.41	24.98
std.dev	3.62	8.9

Examples of finished evaluations

[FIG. 2]

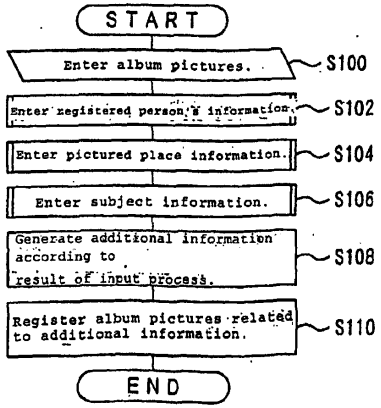


FIG. 14

FIG. 14A

FIG. 14B

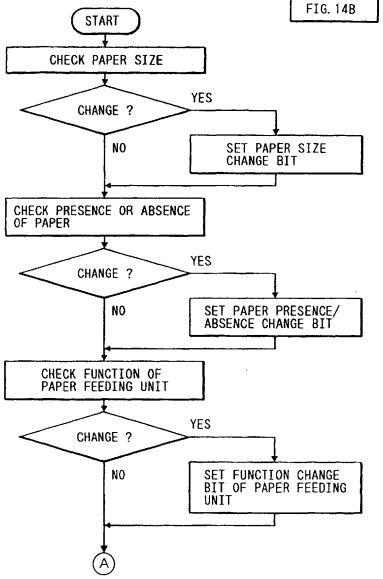


FIG.2

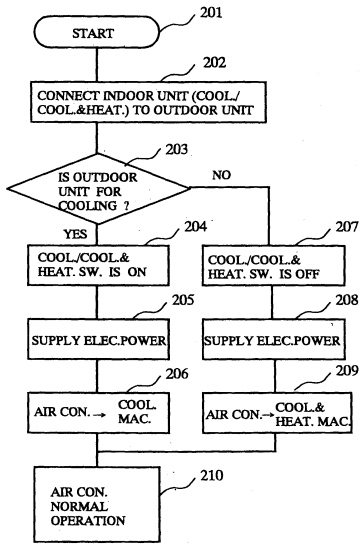
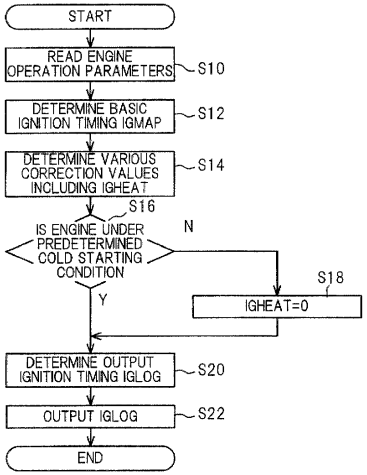


FIG.2



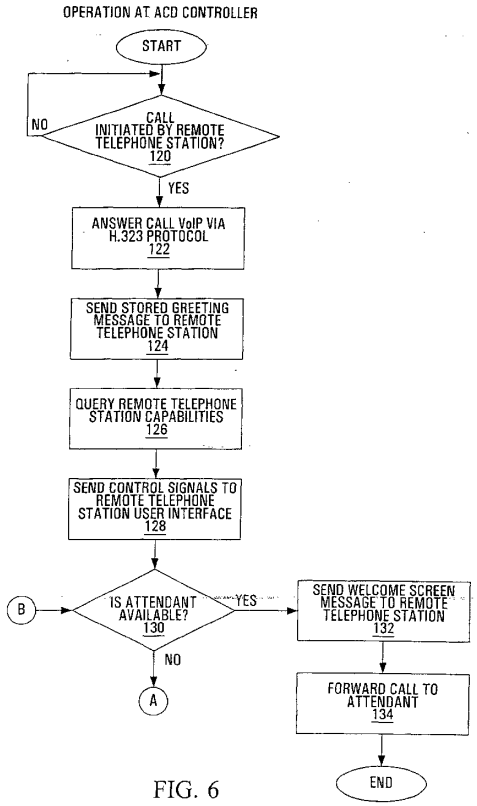


FIG. 6

Unfinished evaluations

FIG. 32

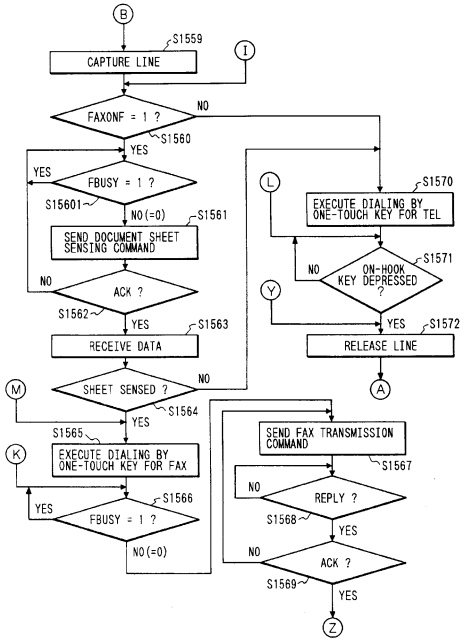
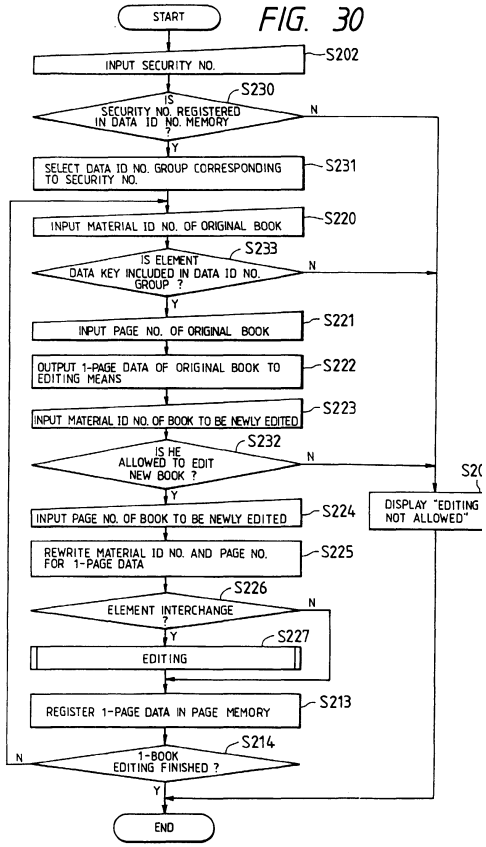


FIG. 30



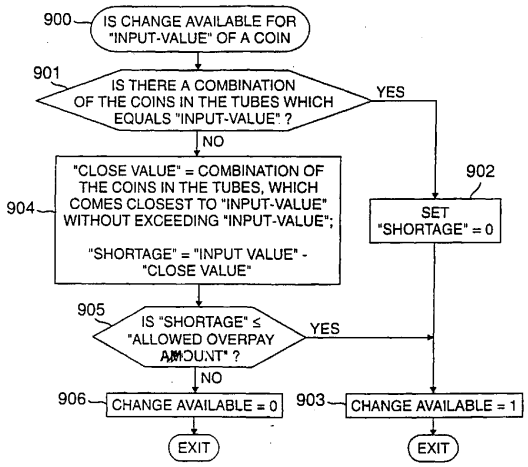


FIG. 8

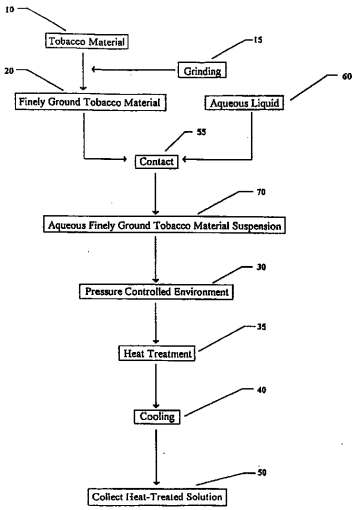


FIG. 1

Measures

- Idea / Motivation
  - The process depicted is important
    - Links between nodes
      - Graph structure
        - » Most Common Sub-graph
  - Node types are less important
    - Only evaluate after structure recognition
  - Text is important
    - Provided that the structure is recognized

Measures (cont.)

- Most common sub-graph (MCS)
  - It's the largest sub-graph common to all graphs in a set of graphs
  - McGregor algorithm
    - Backtracking
    - High computational costs
  - Modified to find all variants of the most common sub-graph
    - Because (we think) filtering on node type would be too restrictive
    - Even more complex

Measures (cont.)

- Node type match & Edit distance
  - Taken separately as the best of all different variants of largest common sub-graphs
  - E.g. if for topic X, 5 different ways to match nodes of run Y were found (all having a score of 0.7 in the MCS), compute the node type match & edit distance for each
    - Node-type match : 0.5 0.5 0.9 0.9 0.5
    - Edit distance: 100 100 90 120 90
    - Then the result scores are:
      - Node match: 0.9
      - Edit distance: 90

Edit Distance

- Smaller is better
  - Smaller?
    - no nodes were returned
    - Actually good match
  - Only comparable in relation with MCS

What’s up next?

- But first…

Information Retrieval Special Issue on IR in the Intellectual Property Domain

- Goal: present cutting-edge research results on open topics related to IR in the Intellectual Property domain in order to advance the current state-of-the-art
- Submissions encouraged to make use of evaluation campaign datasets: CLEF-IP 2011/2012, TREC-CHEM 2011 and NTCIR

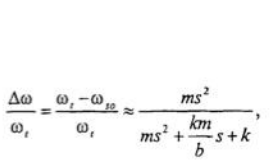
- Submissions due: **15 March 2013**
- More info: https://sites.google.com/site/sipatentir

What’s up next?

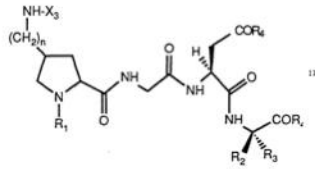
- Overview and plans

Image retrieval in patents

- Must be approached by image type
- Step 1:
  - classification

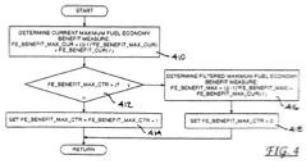


(a) Mathematical

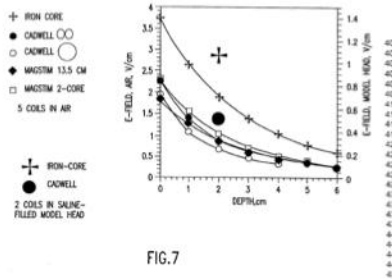


(b) Chemical Formula

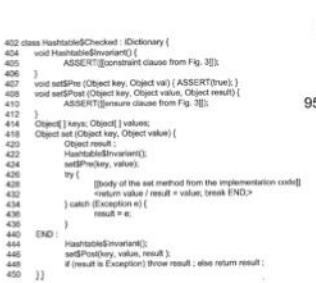
(c) Genetic Sequence



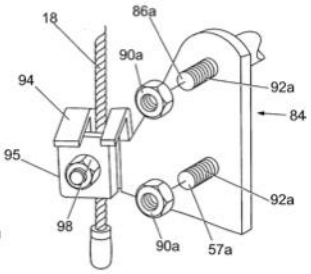
(d) Flow-chart



(e) Graph

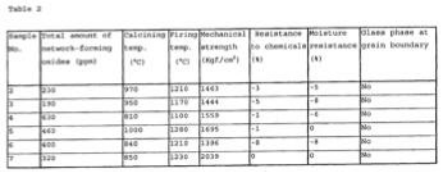


(f) Program Listing



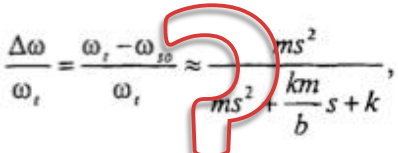
(g) Abstract Drawing

(h) Symbol

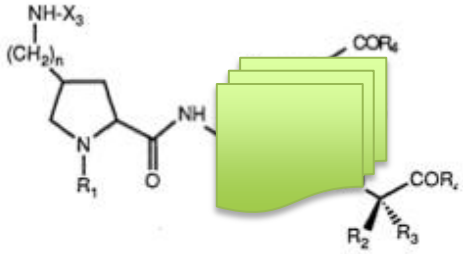


(i) Table

Figure 1: Examples of types of figures in patents

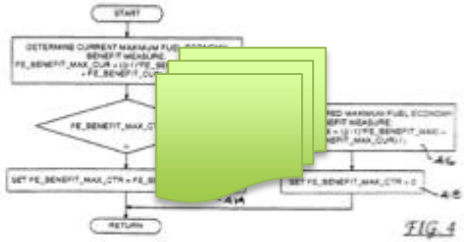


(a) Mathematical



(b) Chemical Formula

(c) Genetic Sequence



(d) Flow-chart

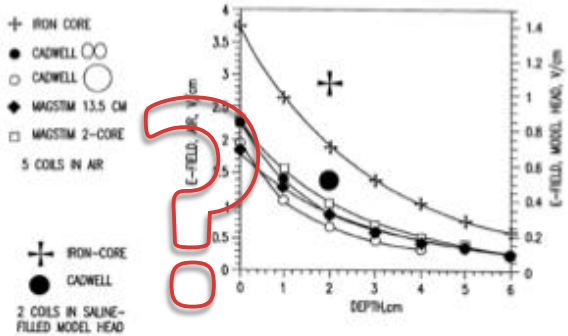


FIG.7

(e) Graph

```
402 class Hashtable$Checked : IDictionary {
404     void Hashtable$invariant() {
405         ASSERT([[constraint clause from Fig. 3]]);
406     }
407     void set$Pre (Object key, Object value) { ASSERT(true); }
408     void set$Post (Object key, Object value, Object result) {
410         ASSERT([[ensure clause from Fig. 3]]);
412     }
414     Object[ ] keys; Object[ ] values;
416     Object set (Object key, Object value) {
420         Object result;
422         Hashtable$Invariant();
424         set$Pre(key, value);
426         try {
428             [[body of the set method from the implementation code]]
432             <return value / result = value; break END;>
434         } catch (Exception e) {
436             result = e;
438         }
440     END:
444         Hashtable$invariant();
446         set$Post(key, value, result);
448         if (result is Exception) throw result ; else return result ;
450     }}
```

(f) Program Listing

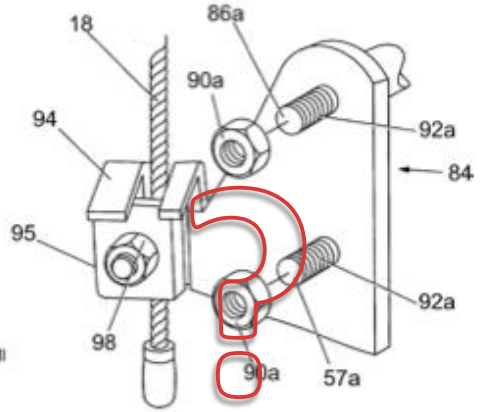


Fig. 10

(g) Abstract Drawing

(h) Symbol

Table 2

Sample No.	Total amount of network-forming oxides (ppm)	Calcining temp. (°C)	Firing temp. (°C)	Mechanical strength (kgf/cm²)	Resistance to chemicals (h)	Moisture resistance (h)	Glass phase at grain boundary
2	230	970	1210	1[illegible]	-1	-5	No
3	190	950	1170	1[illegible]44	-5	-8	No
4	630	810	1100	1[illegible]59	-1	-6	No
5	440	1000	1200	1693	-1	0	No
6	400	840	1210	1396	-8	-8	No
7	320	850	1230	2039	0	0	No

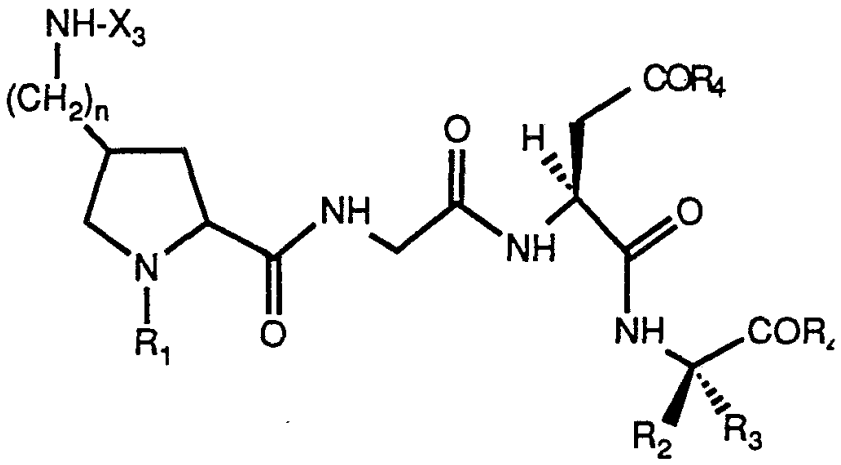
(i) Table

Figure 1: Examples of types of figures in patents

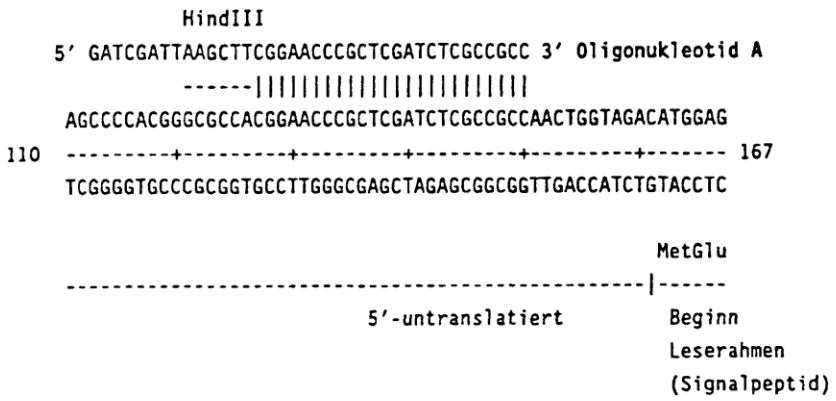
Mathematical formulas

$$\frac{\Delta\omega}{\omega_t} = \frac{\omega_t - \omega_{so}}{\omega_t} \approx \frac{ms^2}{ms^2 + \frac{km}{b}s + k},$$

Chemistry



Genetic data



Plots/Graphs

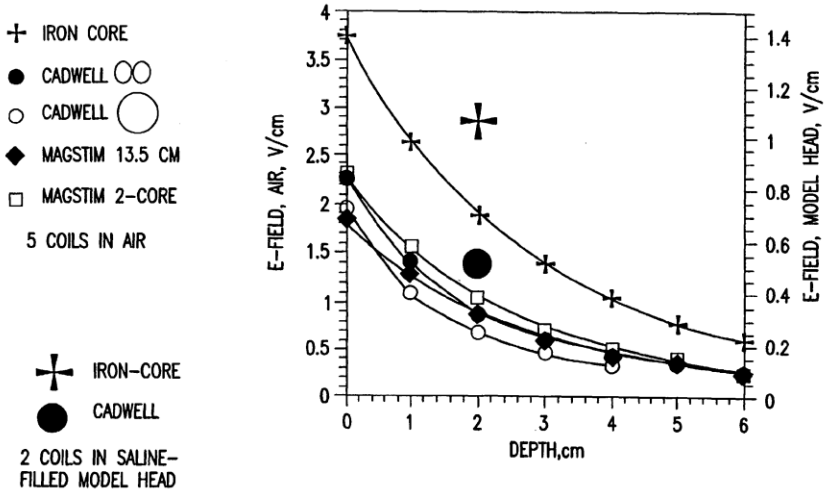


FIG.7

Code listings

```
402 class Hashtable$Checked : IDictionary {
404     void Hashtable$Invariant() {
405             ASSERT([[constraint clause from Fig. 3]]);
406     }
407     void set$Pre (Object key, Object val) { ASSERT(true); }
408     void set$Post (Object key, Object value, Object result) {
410             ASSERT([[ensure clause from Fig. 3]]);
412     }
414     Object[ ] keys; Object[ ] values;
418     Object set (Object key, Object value) {
420             Object result ;
422             Hashtable$Invariant();
424             set$Pre(key, value);
426             try {
428                     [[body of the set method from the implementation code]]
432                     <return value / result = value; break END;>
434             } catch (Exception e) {
436                     result = e;
438             }
440     END :
444             Hashtable$Invariant();
446             set$Post(key, value, result );
448             if (result is Exception) throw result ; else return result ;
450     } }
```

Drawings

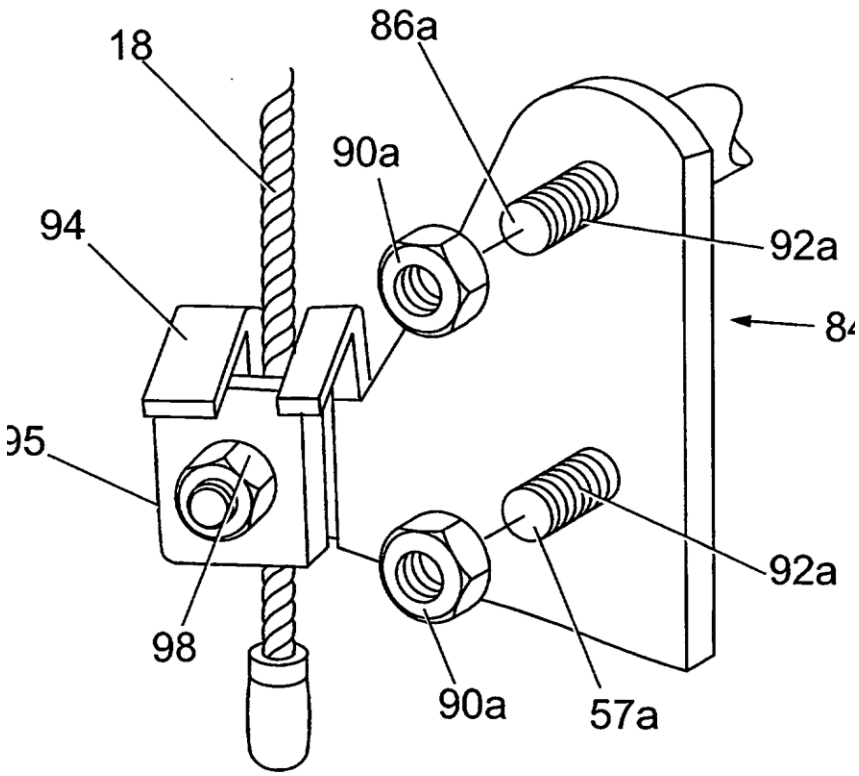
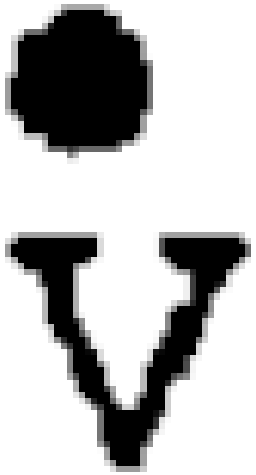


Fig. 10

Characters/symbols



Tables

Table 2

Sample No.	Total amount of network-forming oxides (ppm)	Calcining temp. (°C)	Firing temp. (°C)	Mechanical strength (Kgf/cm^2)	Resistance to chemicals (%)	Moisture resistance (%)	Glass phase at grain boundary
2	230	970	1210	1463	-3	-5	No
3	190	950	1170	1444	-5	-8	No
4	630	810	1100	1559	-1	-6	No
5	460	1000	1280	1695	-1	0	No
6	400	840	1210	1396	-8	-8	No
7	320	850	1230	2039	0	0	No