

## PubMed からのテキストマイニングによる疾患と遺伝子の関連のデータベース LEGENDA

膨大な量の医学・生物学分野の文献を有効活用するためには、テキストマイニング技術を使った機械的な知識抽出が必要である。そこでわれわれは、PubMed から機械的に遺伝子や疾患等の関連を抽出し、その結果をまとめたデータベース LEGENDA (Literature-Extracted GENE-Disease Associations) を構築した。まず、ヒトの遺伝子名、疾患名、遺伝子機能名、化合物名をあらゆる用語を集め、同義語をまとめて分類した。そして、2009年11月時点のPubMedに登録されている約1890万件の文献要旨の中で、これらの用語が同じ文中に出現する(共起する)文を探した。その結果として得られた共起文をデータベースに登録し、検索機能を整備して、LEGENDA ver. 3 を構築した(近日公開予定)。

LEGENDA では、ユーザがある用語を入力すると、それと PubMed 中で共起する他の用語の一覧がスコア順に表示される。さらに、用語が共起する文献のタイトルや要旨を表示したり、ダウンロードしたりすることができる。また、LEGENDA では相互情報量に基づくスコアを使って概念どうしの関連の強さを評価している。このスコアを用いて、直接共起だけでなく、他の概念を仲介することにより関連が推定される場合(間接共起)の探索を行う機能も実現した。

LEGENDA は、<http://hin.vj.jp/legenda/> で利用できる。

The screenshot shows the LEGENDA web interface. At the top, there is a search bar and a 'Search Relations' button. Below this, the 'Document View' section displays a list of documents with co-occurrences of HBB and Thalassemia. The table has three columns: PMID, Published Date, and Title. The first few rows are as follows:

PMID	Published Date	Title
17003927	2006 Oct	Prevalence of thromboembolic events among 8,860 patients with thalassaemia major and intermedia in the Mediterranean area and Iran.
16645164	2006 Aug 15	Flanking HS-62.5 and 3' HS1, and regions upstream of the LCR, are not required for beta-globin transcription.
16886697	2006 Jul 15	[Acute anaemia in a Vietnamese patient with alpha-thalassaemia and a parvovirus infection]
16829478	2006 Jul-Aug	[Beta(o)/beta(o) thalassemia with a mild phenotype]
16480700	2006 Jun	Molecular detection of Spanish deltabeta-thalassemia associated with beta-thalassemia identified during prenatal diagnosis.
16785121	2006 Jun	First report on the co-inheritance of (beta) IVS I-1 (G->T) Thalassemia with the (gamma) CD85 [Phe->Ser (F1) (TTT->TCT)] HbA2 Etoia in Iran.
16628732	2006 May	Hb Florida: a novel elongated C-terminal beta-globin variant causing dominant beta-thalassemia phenotype.
16421096	2006 Mar 17	Humanized beta-thalassemia mouse model containing the common IVSI-110 splicing mutation.
16360115	2006 Mar	Comparison of basic peptides- and lipid-based strategies for the delivery of splice correcting oligonucleotides.
16542675	2006 Feb-Apr	Erythrocyte morphology and haemoglobin types of neonatal roan antelopes (Hippotragus equinus) with hypochromic poikilocytic anaemia.

At the bottom of the table, there is a 'Page 1 of 13' indicator and a 'Show abstract' button. The footer of the page reads 'Copyright©2005-2009 JBIC, AIST. All Rights Reserved' and features the AIST logo.

図. LEGENDA の共起文献の一覧表示画面。この例は beta-globin と thalassemia が共起する文を含む論文の一覧である。上部テキストボックスに遺伝子名、疾患名、化合物名等を入れて共起する用語の組を選択すると、この画面が表示される。この画面では PubMed の要旨を表示したりダウンロードしたりすることもできる。