

Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study

Kristina M. Hettne^{1,2}, Marc Weeber^{2,3},
Marja L. Laine⁴, Hugo ten Cate⁵,
Scott Boyer¹, Jan A. Kors² and
Bruno G. Loos⁴

¹Safety Assessment, AstraZeneca R&D Mölndal, Sweden; ²Department of Medical Informatics, Erasmus MC, Rotterdam, The Netherlands; ³Knewco Inc., Rockville, MD, USA; ⁴Department of Periodontology, Academic Center for Dentistry Amsterdam (ACTA), University of Amsterdam and Vrije University, The Netherlands; ⁵Department of Internal Medicine and Cardiovascular Research Institute Maastricht, University Hospital Maastricht, The Netherlands

Hettne KM, Weeber M, Laine ML, Cate Ht, Boyer S, Kors JA, Loos BG. Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study. *J Clin Periodontol* 2007; 34: 1016–1024. doi: 10.1111/j.1600-051X.2007.01152.x.

Abstract

Aim: The aim of the current report was to generate and explore new hypotheses into how, in a pathophysiological sense, atherosclerosis and periodontitis could be linked.

Material and Methods: Two different biomedical informatics techniques were used: an association-based technique that generated a ranked list of genes associated with the diseases, and a natural language processing tool that extracted the relationships between the retrieved genes and lipopolysaccharide (LPS).

Results: This combined approach of association-based and natural language processing-based literature mining identified a hit list of 16 candidate genes, with PON1 as the primary candidate.

Conclusions: Further study of the literature prompted the hypothesis that PON1 might connect periodontitis with atherosclerosis in both an LPS-dependent and a non-LPS-dependent manner. Furthermore, the resulting genes not only confirmed already known associations between the two diseases, but also provided genes or gene products that have only been investigated separately in the two disease states, and genes or gene products previously reported to be involved in atherosclerosis. These findings remain to be investigated through clinical studies. This example of multidisciplinary research illustrates how collaborative efforts of investigators from different fields of expertise can result in the discovery of new hypotheses.

Key words: atherosclerosis; biomedical informatics; information storage and retrieval; lipopolysaccharides; periodontitis

Accepted for publication 17 August 2007

Conflict of interest and source of funding statement

The authors declare that they have no conflict of interests.

The work reported in this study was sponsored by the EC funded Network of Excellence INFOBIOMED (IST-507585).

Increasing evidence is appearing in the literature that cardiovascular diseases (CVDs) and periodontitis are associated and possibly causally linked (Janket et al. 2003, Cueto et al. 2005, Soder et al. 2005, Spahr et al. 2006). Periodontitis is a chronic infectious disease of the supporting tissues of the teeth. The

aetiology of the disease is strongly related to the colonization of the periodontal tissues by a complex mix of anaerobic, Gram-negative bacteria. However, Gram-positive bacteria also constitute a significant component of subgingival biofilm (Pihlstrom et al. 2005). The ensuing inflammatory reaction is responsible for

the slow progression of loss of periodontal ligament and alveolar bone; if left untreated, teeth become mobile and eventually exfoliate (Pihlstrom et al. 2005).

The underlying pathology of atherosclerotic CVD is the chronic formation and extension of atherosclerotic lesions in blood vessels. It is generally appreciated that atherosclerosis has a marked inflammatory component and might be considered an inflammatory disease (Ross 1999, Mullenix et al. 2005). Moreover, bacteria may influence atherosclerotic plaque progression either by colonization or perhaps more likely by leaving a "fingerprint" due to infection of plaque invading leukocytes (Ott et al. 2006). Either way, bacteria may precipitate an increased inflammatory state in the vessel walls, predisposing for plaque rupture and thrombosis formation (Cairo et al. 2004).

The current hypothesis on the epidemiological association between CVD and periodontitis deals with the notion that in periodontitis the ulcerated pocket epithelium forms a porte d'entrée for microorganisms to enter the circulation and distant organs (Loos 2005). Indeed, transient low-grade bacteraemias occur in chronic periodontitis, possibly daily, for example during chewing and tooth brushing (Kinane et al. 2005), and bacteria from the oral cavity and other bodily locations have been identified in atherosclerotic plaques (Fiehn et al. 2005). Moreover, a direct relationship between periodontal microbiology and subclinical atherosclerotic lesions has been observed (Desvarieux et al. 2005). Thus it is thought that bacteria and/or bacterial components such as endotoxin might exacerbate inflammatory processes in atherosclerotic lesions (Tabrizi et al. 2007), resulting in the increased chance for fibrous plaque rupture and thrombus formation, with consequently ischemia in the distal tissues.

For both CVD and periodontitis, genetic susceptibility is widely accepted (Loos et al. 2005, Nordlie et al. 2005, Yoshie et al. 2007). In fact, a common genetic background for atherosclerosis and periodontitis has been suggested (Kornman & Duff 2001, Nichols et al. 2001, Mattila et al. 2005). Especially, both diseases might be linked due to a genetically determined hyper-responsive inflammation trait or a genetically determined lack of appropriate control of the immune reactions to bacteria and/or bacterial components such as endotoxin (lipopolysaccharide or "LPS") (Chun et al. 2005).

In addition to clinical and pathophysiological research, the use of biomedical informatics (BMI) is another avenue into exploring possible mechanisms by which complex diseases are interrelated. The science of BMI applies informatics techniques to organize bio-molecular data on a large scale and relate it to clinical practice (Luscombe et al. 2001). The combination of medical and bioinformatics approaches is expected to result in significant advantages in both understanding mechanisms of disease and inter-individual susceptibility, which in turn will open new possibilities in individualized medical health care (Kohane 2000). Information retrieval tools developed for analyzing biomedical literature can be helpful in systemizing the existing knowledge and in displaying it in such a way that relations can become visible (Shatkay 2005), and they have been proven useful for literature-based hypothesis generation (e.g. Smalheiser & Swanson 1998, Weeber et al. 2003). Two main approaches to automatically retrieve and extract information from biomedical literature can be distinguished: (1) computation of the co-occurrence of biomedical concepts and (2) extraction of the actual relations between biomedical concepts using natural language processing. In this report, we started with a variation of the co-occurrence-based approach where we, similar to (Glenisson et al. 2004), rather than using direct co-occurrence, create concept vectors representing the textual context surrounding a concept. The advantage of using this approach is that two concepts do not actually have to co-occur in any given abstract of a journal article for an association to be created. Instead, the genes can be associated because they share many common concepts; the more the common concepts, the higher the association score. We used this association-based approach to generate a ranked hit list of genes associated with the diseases periodontitis and atherosclerosis, after which the exact relationships between the retrieved genes were identified using a natural language processing-based tool in order to elucidate previously unknown connections between the diseases. The common component of Gram-negative periodontal pathogens is LPS, and thus LPS might constitute the trigger for events related to a possible link between periodontitis and atherosclerosis. In order to provide a disease context for the genes, LPS was included

in the model. The results were thereafter analyzed and judged for relevance by domain experts (BL, HC). This approach exemplifies the benefit of using BMI tools to assist the clinician and the basic scientist and their complementary expertise to advance the understanding of potential relationships between complex diseases and generate new hypotheses for further research.

Material and Methods

Creating a concept profile of periodontitis and atherosclerosis from textual data

The initial step in generating a ranked hit list of genes associated with periodontitis and atherosclerosis was to produce a concept profile for both diseases. First, two separate sets of MEDLINE (MEDLINE 2007) abstracts annotated with Medical Subject Headings (MeSH) (Lipscomb 2000) terms for either periodontitis or atherosclerosis were retrieved (1995–2005). MeSH headings are manually assigned to abstracts by indexers at the National Library of Medicine (NLM 2007). The MeSH headings were used to retrieve the abstracts as they provide important metadata for a document: they describe what the document is about in a consistent way, thus reducing the need for term variation when the search is conducted in the MEDLINE database. For example, the MeSH term "periodontitis" will be assigned to a document if a document is judged by the indexer to be about periodontitis even if a synonym for periodontitis is being used in the document instead of the MeSH term. Using Collexis technology (van Mulligen et al. 2000, Collexis 2007), the documents were indexed with two thesauri: MeSH and the Gene Ontology (Ashburner et al. 2000). For every document, this resulted in a list of concepts that were rank-ordered according to their Term Frequency \times Inverse Document Frequency (TF-IDF) scores, as a measure of their relevance (Gerard & Christopher 1988). Each document has thus been transformed into a document profile. For each disease, all document profiles containing the disease were then aggregated into one concept profile by averaging the rank of all concepts over all profiles, thereby creating one extended disease profile for periodontitis and one for atherosclerosis. In this way, a document-centered profile has been transformed into a

disease-centered profile. The two separate disease profiles were then aggregated into one periodontitis–atherosclerosis profile by averaging over concepts again.

Matching the periodontitis–atherosclerosis profile with gene concept profiles

The next step was to produce gene profiles that could be matched with the periodontitis–atherosclerosis profile. The National Centre for Biotechnology Information (NCBI) EntrezGene (Maglott et al. 2005, EntrezGene 2007) April 2005 database dump was used as a source of information on genes. For each human, mouse, and rat gene, all available annotation information, PubMed identifiers, and Gene References Into Function (GeneRIFs) were extracted. The concept profiles for the genes in the Entrez Gene database were created in a similar fashion as the disease fingerprint, again using Collexis technology.

The periodontitis–atherosclerosis profile was thereafter matched with each gene profile and rank-ordered using a cosine vector-matching score (Salton 1989). The result of the match is a list of genes ranked by contextual similarity. Due to the explorative nature of the present study, we limited ourselves to the resulting list of genes with at least a 30% contextual similarity, where it has to be understood that the best match is set at 100%.

Exploring relationships between genes in the hit list – interaction with LPS

Assuming bacteraemia is occurring in periodontitis, the presence of bacteria/bacterial components could trigger a systemic inflammatory response, which favours an atherosclerotic process (Nichols et al. 2001). A common component of Gram-negative periodontal pathogens is endotoxin, i.e. LPS, which may constitute the trigger for events related to a possible link between periodontitis and atherosclerosis. We used PathwayStudio (PathwayStudio 2007), a software application developed for navigation and analysis of biological pathways, gene regulation networks and protein interaction maps to explore how the previously identified genes interconnect with LPS in a pathophysiological model of atherosclerosis. The genes from the hit list with $\geq 30\%$ similarity were entered into PathwayS-

tudio and the option “find only direct interactions between selected nodes” was used to build a network. PathwayStudio has been described in more detail by Nikitin et al. (2003) and PathwayStudio (2007). Briefly, PathwayStudio finds connections between concepts, called “nodes”, by searching through a database of interactions derived from the literature by the natural language processing software MedScan (Novichkova et al. 2003). MedScan performs a full grammatical and semantic analysis of MEDLINE (Novichkova et al. 2003, Daraselia et al. 2004). When parsing a MEDLINE document, a semantic interpreter of the natural language processing component transforms the syntactic structure into a semantic structure. The output of the semantic parse is the input for an ontological analysis that was developed by Daraselia et al. (2004), in which an “entity” is represented as a protein, a cellular object, a cellular process or a small molecule, and “controls” describe functional relationships between these entities. Relations between entities are stored in a relational database, and these relations can then be displayed and explored through a graphical interface. The underlying facts that contributed to the relation between two entities are provided as complete references and links out to PubMed (PubMed 2007).

The default option of the direct interaction algorithm in PathwayStudio only detects direct relationships between two nodes; mechanisms that require a new intermediate node (a node not present in the original list of nodes that was provided as input) would not be detected. Such intermediate nodes, however, might represent knowledge that the scientist entering the genes was not aware of at the time of input, such as factors acting as enhancers/modifiers for a critical step in a pathway, or co-factors needed in order for a transcription factor to bind to DNA. The direct interaction algorithm therefore also provides an option to include these kinds of intermediate nodes; such an intermediate node allows the creation of a triplet that connects two input nodes. We used this option to create a second network, where we could find intermediate nodes between the original nodes in our first network. When a new node is added to the network, all links, to and from that node to all other nodes in the network, are displayed.

Results

Genes identified by the association-based method

A ranked list of genes associated with the diseases periodontitis and atherosclerosis was generated. The gene paraoxonase 1 (PON1) was identified as having the strongest influence in the relationship between atherosclerosis and periodontitis and was therefore set by default at 100% similarity. All the other genes in Table 1 have similarity percentages relative to PON1. We concentrated on genes with at least 30% similarity; in this way 16 genes were identified (Table 1). By manual investigation of available literature using advanced PubMed queries containing both MeSH terms and synonyms not provided by MeSH, three distinguished groups emerged:

1. Seven genes or gene products that previously have been suggested to have a similar function in both diseases: CD14 molecule (CD14) (Hajishengallis et al. 2002, 2004), toll-like receptor 2 (TLR2) (Hajishengallis et al. 2002, 2004), adiponectin, C1Q and collagen domain containing (ADIPOQ) (Iwamoto et al. 2003), toll-like receptor 4 (TLR4) (Hajishengallis et al. 2002, 2004), 5,10-methylenetetrahydrofolate reductase (MTHFR) (Kornman & Duff 2001), Fc fragment of IgG, low affinity IIa, receptor (FCGR2A) (Naito et al. 2006), apolipoprotein E (APOE) (Li et al. 2002, Lalla et al. 2003, Gibson et al. 2004).
2. Three genes or gene products that have been separately investigated in both diseases, but never before suggested to play a connecting role: tumor necrosis factor receptor superfamily, member 11b (TNFRSF11B) (Garlet et al. 2003, Rothenbacher et al. 2006), chemokine (C-X-C motif) ligand 10 (CXCL10) (Silva et al. 2005, Schoppet et al. 2006), chemokine (C-C motif) receptor 2 (CCR2) (Taubman et al. 2005, Veillard et al. 2005).
3. Six genes or gene products that have only been suggested to play a role in atherosclerosis and not periodontitis: paraoxonase 1 (PON1) (Ng et al. 2005, Rodriguez Esparragon et al. 2006), paraoxonase 2 (PON2) (Ng et al. 2006), oxidised low density lipoprotein (lectin-like) receptor 1

Table 1. Ranked list of genes with at least 30% similarity between periodontitis and atherosclerosis relative to the highest similarity (set at 100%), identified using an association-based method

Gene symbol	Full name	Periodontitis and atherosclerosis	Periodontitis	Atherosclerosis	Percent similarity	Group
PON1	Paraoxonase 1	No	No	Yes	100	3
CD14	CD14 molecule	Yes	Yes	Yes	85	1
TLR2	Toll-like receptor 2	Yes	Yes	Yes	71	1
ADIPOQ	Adiponectin, C1Q and collagen domain containing	Yes	Yes	Yes	66	1
PON2	Paraoxonase 2	No	No	Yes	56	3
TLR4	Toll-like receptor 4	Yes	Yes	Yes	48	1
OLR1	Oxidised low density lipoprotein (lectin-like) receptor 1	No	No	Yes	48	
MTHFR	5,10-Methylenetetrahydrofolate reductase	Yes	Yes	Yes	47	1
CXCL16	Chemokine (C-X-C motif) ligand 16	No	No	Yes	45	3
SCARB1	Scavenger receptor class B, member 1	No	No	Yes	43	3
TNFRSF11B	Tumor necrosis factor receptor superfamily, member 11b	No	Yes	Yes	40	2
FCGR2A	Fc fragment of IgG, low affinity IIa, receptor	Yes	Yes	Yes	40	1
CXCL10	Chemokine (C-X-C motif) ligand 10	No	Yes	Yes	36	2
ADD1	Adducin 1 alpha	No	No	Yes	32	3
CCR2	Chemokine (C-C motif) receptor 2	No	Yes	Yes	32	2
APOE	Apolipoprotein E	Yes	Yes	Yes	30	1

Official gene symbols and full names are used (HUGO Gene Nomenclature Committee 2007, Wain et al. 2002). Advanced PubMed searches were performed to find documents that mention a gene together with each separate disease or with both diseases combined. If no such documents were found, the corresponding entry contains No, otherwise Yes. The group column refers to the groupings made based on these advanced PubMed searches.

(OLR1) (Vohra et al. 2006), chemokine (C-X-C motif) ligand 16 (CXCL16) (Sheikine et al. 2005), scavenger receptor class B, member 1 (SCARB1) (Rodriguez Esparragon et al. 2006), adducin 1 alpha (ADD1) (van Rijn et al. 2006).

No genes were found by this method to have only been linked to periodontitis and not to atherosclerosis.

Interactions between LPS and the genes identified by the association-based method

The entity-relationship tool PathwayStudio was used to find direct interactions between genes from the association-based analysis and LPS; it was found that LPS directly regulates nine of the rank-listed genes (CXCL10, OLR1, CCR2, CD14, TLR2, TLR4, FCGR2A, SCARB1, and APOE) (Fig. 1). The genes for which PathwayStudio could not find a direct link to LPS were PON1, PON2, MTHFR, TNFRSF11B, ADD1, ADIPOQ, and CXCL16. To confirm that no interaction with LPS could be found in the literature between LPS and the genes mentioned, a manual advanced search in PubMed was conducted. The search revealed that PON1, TNFRSF11B, ADIPOQ, and CXCL16 are regulated by LPS (bin Ali et al. 2003, Xu et al. 2005, Chung et al. 2006, Peake et al. 2006). This leaves

PON2, ADD1, and MTHFR for which no link with LPS could be found. However, a structurally similar molecule to LPS called lipoarabinomannan has been shown to induce an increase in ADD1 (encoded by the ADD1 gene) protein phosphorylation (Hestvik et al. 2003). The next step was to let the direct interaction algorithm in PathwayStudio search for intermediate nodes between the original nodes in our first network. The resulting network, which allowed new nodes to be incorporated, is shown in Fig. 2. In addition to the 16 original rank-listed genes that were provided as input nodes, three new nodes (i.e. nodes not in the original list of 16 genes) have been added to the model: interleukin 8 (IL8), LPS-binding protein (LBP), and signal transducer and activator of transcription 1, 91 kDa (STAT1). LPS directly regulates IL8, LBP, and STAT1 in the model (Fig. 2). IL8 has been explored previously as a connecting concept between periodontitis and atherosclerosis by, for example, Fokkema et al. (2003) and can thus be added to group 1. LBP has been investigated in both diseases separately (Dunzendorfer et al. 2004, Ren et al. 2004), but has not previously been suggested to link the two diseases, which qualifies it for group 2. STAT1 has been shown to be part of the vascular smooth muscle cell apoptosis pathway, which occurs in atherosclerotic plaques, contributing to plaque instability (Rosner

et al. 2006), but has not been mentioned together with periodontitis in the literature, which qualifies it for group 3.

Discussion

This study was initiated to employ new techniques in BMI for hypothesis generation concerning the epidemiological link between periodontitis and atherosclerosis.

Relationships between the hit list of genes and the two diseases

Among the 16 genes in the hit list using the association-based method, the PON1 gene was identified as having the highest association with atherosclerosis and periodontitis. Epidemiological, genetic, and biochemical studies have brought forward that PON1 encodes for a protein that has a protective effect against atherosclerosis, although the exact mechanism behind the effect is unclear (Ng et al. 2005, Ng et al. 2006). An advanced PubMed search did not retrieve any documents mentioning PON1 in relation to periodontitis, but theoretically it may be associated with this condition. One study reported that PON1 activity is regulated at the mRNA level in a gender-specific manner by LPS (bin Ali et al. 2003), the common component of many periodontal bacteria. Moreover, Ozer et al.

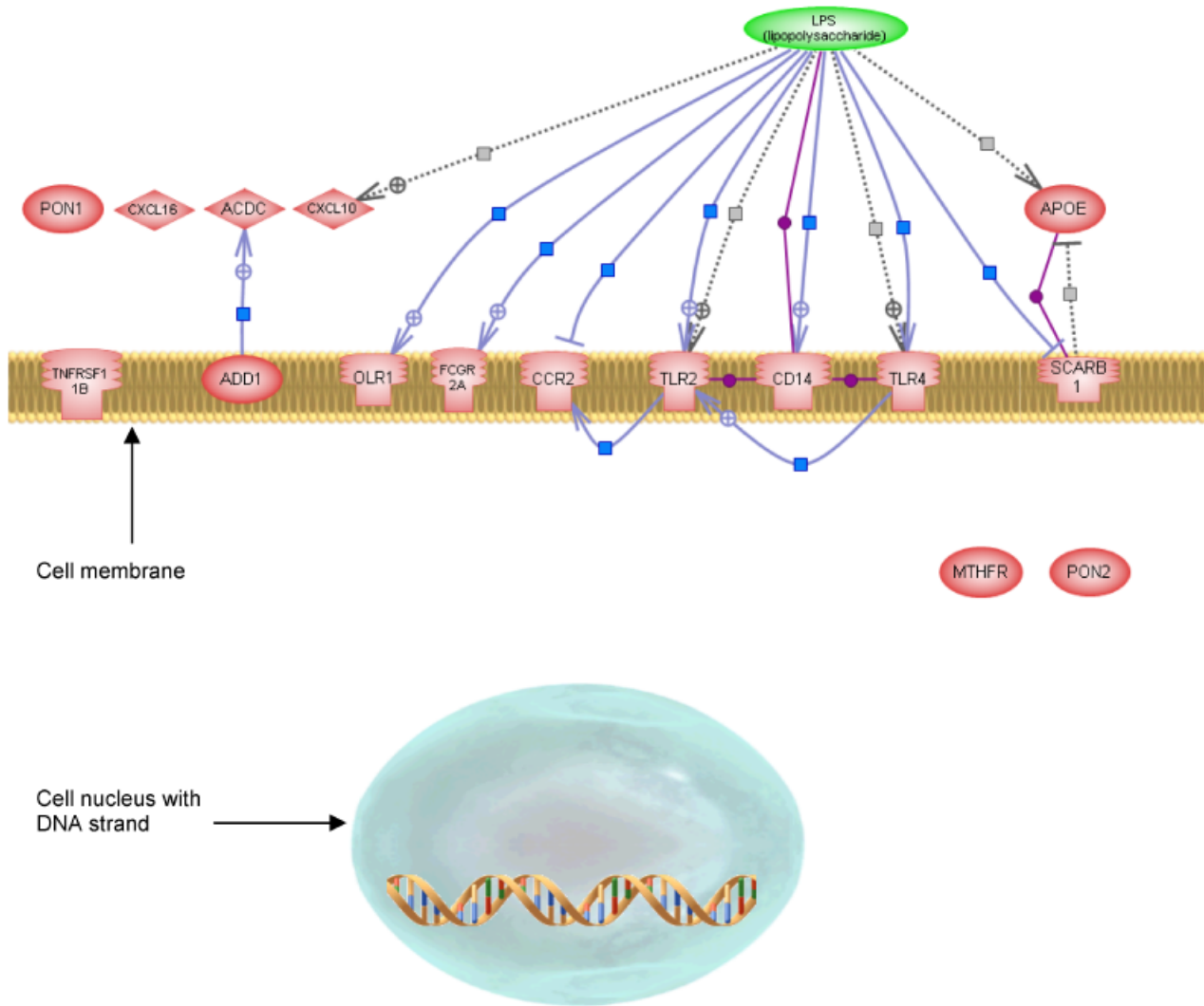


Fig. 1. Direct links found in PathwayStudio (Nikitin et al. 2003, PathwayStudio 2007) between the entered genes and lipopolysaccharide. Red circles denote proteins, red diamond shapes denote ligands, red rattle shapes denote receptors, and green circles denote small molecules. Grey dotted lines denote regulation, blue lines denote expression, and purple lines denote binding.

(2005) demonstrated that paraoxonase-induced interruption of bacterial communication represents a novel mechanism to modulate quorum sensing by bacteria; this could possibly take place both in biofilms of periodontal infections and in biofilms in atherosclerotic plaques. However, the consequences for host immunity are yet to be determined (Ozer et al. 2005). While a low PON1 activity is associated with CVD (low inhibition of LDL oxidation), its role in inflammation is only recently suggested. In patients with systemic lupus erythematosus (SLE) also a lower PON1 activity was found in comparison with controls; genetic polymorphisms in the PON1 gene were associated with this lower activity (Tripi et al. 2006). Similar to an increased prevalence of CVD in

patients with SLE, increased CVD in periodontitis cases could be related to the genetic polymorphisms in the PON1 gene as found in SLE. The resulting lower PON1 activity results in increased levels of oxidized LDL, which subsequently promotes foam cell formation and atherosclerotic streaks in the arterial cell walls (Tripi et al. 2006). PON1 might thus connect periodontitis with atherosclerosis in both an LPS-dependent and a non-LPS-dependent manner.

We made a distinction between (1) genes or gene products that previously have been suggested to link periodontitis with atherosclerosis by some common biological mechanism, (2) genes or gene products that have been associated with both diseases separately

but have not before been suggested to link the two diseases, and (3) genes or gene products that previously have only been associated with atherosclerosis. These groupings are interesting in different ways. The first group of genes or gene products, i.e. CD14, TLR2, ADIPOQ, TLR4, MTHFR, FCGR2A, APOE, and IL8, does not provide increased insight into a possible link between the diseases because they have previously been reported as such in the literature, but they add to the validity of the methods used in this paper. The second group of genes or gene products, i.e. TNFRSF11B, CXCL10, CCR2, and LBP, is interesting as they represent two separate knowledge domains (periodontitis and atherosclerosis), between which these genes or

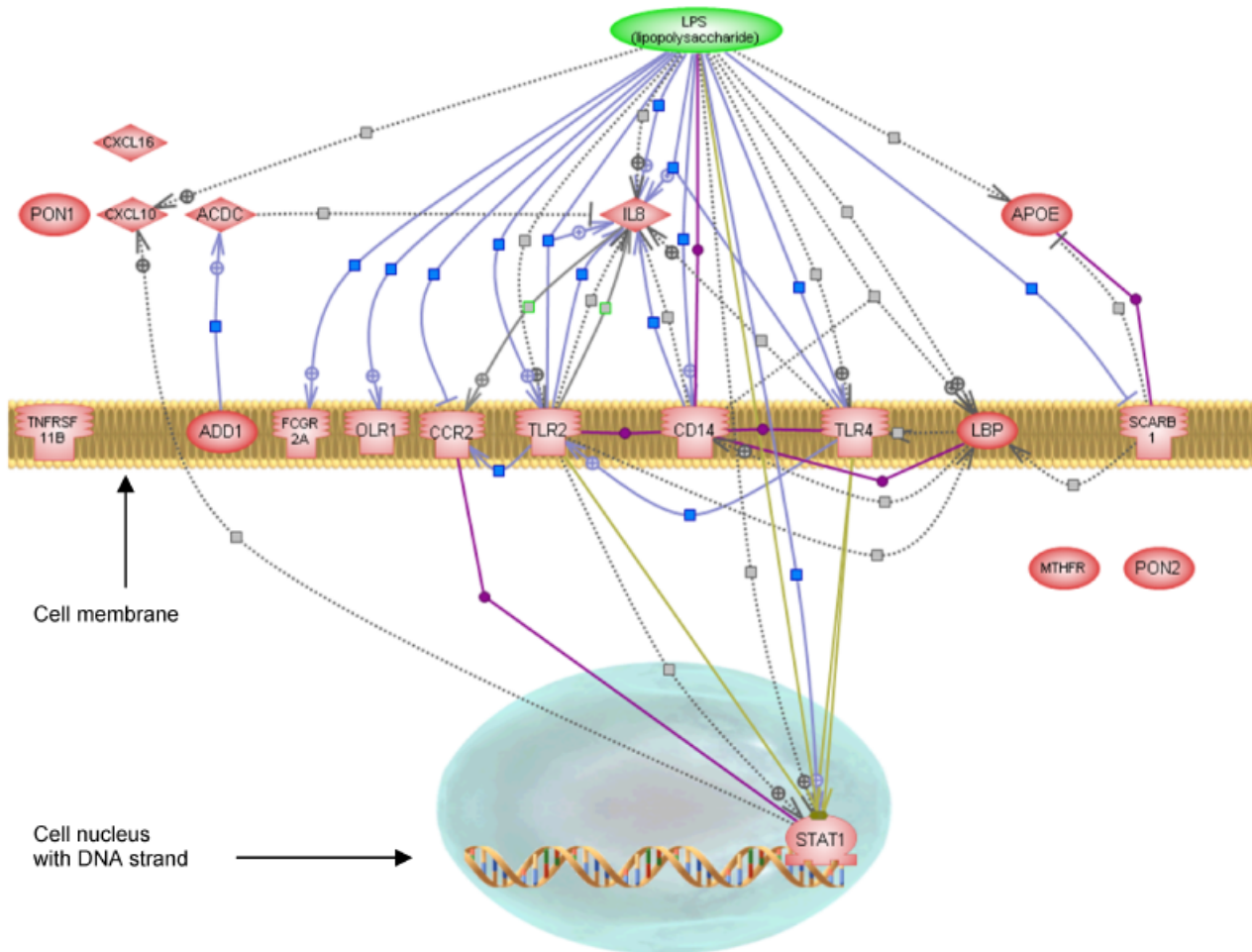


Fig. 2. Links found in PathwayStudio (Nikitin et al. 2003, PathwayStudio 2007) between the entered genes and lipopolysaccharide, allowing for new nodes representing new genes that constitute a three-way connection. Red circles denote proteins, red diamond shapes denote ligands, red rattle shapes denote receptors, and green circles denote small molecules. Grey dotted lines denote regulation, blue lines denote expression, and purple lines denote binding.

gene products might constitute previously unknown links (for example, common genetic variations or common susceptibility pathways). This is a desirable outcome because it is evident that the two knowledge domains are working on the same genes independently from each other and would benefit from cross-collaboration. These links would have been very time-consuming to find without the use of a text-mining method, because they would require a list of all genes or gene products investigated in the two different domains, which can then be searched for overlap that has previously not been reported in the literature. Manual extraction of gene names from review articles might be used for this purpose, but then recent findings might be missed, and there is also a possibility that the review articles used are not complete. The last group of genes or gene products (group

3), i.e. PON1, PON2, OLR1, CXCL16, SCARB1, ADD1, and STAT1, might provide new targets for periodontitis research, because even though they have never been mentioned together with periodontitis in the literature, they came up high on the output of genes from the association-based method. These genes or gene products were found because their concept profiles shared many common concepts with the common concept profile that was created for periodontitis and atherosclerosis. These genes would have been very hard to find using the NLP-based method, because there is no direct link in literature between them and periodontitis. Also, diseases are not included as concepts in PathwayStudio, which means that such experiments cannot be carried out using only PathwayStudio. Small molecules (e.g. LPS) are however included as concepts

in PathwayStudio, and the use of LPS in a model with periodontitis- and atherosclerosis-connected genes to put them in a disease context are discussed in the next section.

No genes or gene products were found that have only been described in periodontitis but not atherosclerosis. This might be due to the cut-off level used (30% contextual similarity). There is currently no gold standard available for where to set the cut-off and this might be an important issue for further research. Another reason might be the method used to generate the concept profiles, such as the thesaurus used for indexing or the weighting of the concepts in the concept profiles. Recently, Jelier et al. (2007) reported on how they created text-derived concept profiles of genes to support assessment of DNA microarray data, and Schuemie et al. (2007) used concept profile technology

together with cross-species homology searches to predict implicit and explicit biologically meaningful functions for a proteomics study of the nucleolus. In these studies the approach was validated by showing its good performance in classification experiments. Both authors used slightly different algorithms to create the actual concept profiles, but the basic methodology is the same as the one used by the authors of this paper. It might be desirable to be able to manually alter the concept profiles before the actual matching procedure takes place. However, the purpose of this paper is to show how a text-mining method automatically can retrieve a hit list of genes that can be used for further in-depth analysis by field experts aided by a BMI tool, and how altering the concept profiles manually at a very early stage in the analysis might provide an undesirable bias.

LPS as a trigger for events linking periodontitis and atherosclerosis

There are many other biological pathways than the pathway triggered by LPS through which the two conditions might be connected, for example heat shock proteins or cell surface fimbriae, but for simplicity in this report we focused on the LPS pathway. Although we realize that the choice of LPS as the main mediator provides a selection bias and limits the results of the analysis, we consider LPS to be a probable key mediator in this crosstalk and its inclusion in the model allows for a demonstration of what can be done using an NLP-based BMI tool to further explore automatically generated hypotheses. In contrast to the association-based method, the NLP-based method has the power to explain the findings in a model together with LPS. However, the NLP-based method needs genes or small molecules as input in order to start the analysis. For this purpose, the list of genes from the association-based method was used. Starting with the first group, i.e. genes or gene products that previously have been suggested to link the two diseases, all except MTHFR were shown to be regulated by LPS in the models generated by PathwayStudio (Figs 1 and 2) or advanced manual PubMed search. Thus, MTHFR might constitute a non-LPS pathway. All of the genes or gene products in the second group, i.e. genes or gene products that

have been investigated in both diseases, but never before been suggested to link them together, were either in the model or by advanced manual PubMed search found to have LPS as a common factor. Among the third group of genes or gene products that have previously been linked to atherosclerosis but not to periodontitis, only PON2 could not be connected to LPS either in the model or by advanced manual PubMed search. Thus the interpretation of the 16 genes in a model together with LPS provides substantially more weight to the concept that the link between periodontitis and atherosclerosis may go through LPS.

All genes or gene products mentioned in this paper could potentially play a role in the possible connection between atherosclerosis and periodontitis. At the same time, one should bear in mind that the current tools in the biomedical domain are not capable of delivering clear-cut hypotheses. Thus, the interpretation of the data and the judgements of its value should always be made by field experts and investigators. Their assessments can lead to experimental studies to test the generated hypotheses by performing *in vitro* studies, animal models, and/or clinical and epidemiological studies.

Conclusions

A text-mining exercise using association-based and natural language-processing techniques was performed to explore the epidemiological association between CVD (*in casu* atherosclerosis) and periodontitis. This approach identified a hit list of 16 candidate genes, with PON1 as the primary candidate. Further study of the literature prompted the hypothesis that PON1 might connect periodontitis with atherosclerosis in both an LPS-dependent and a non-LPS-dependent manner. Furthermore, the results not only confirmed already known associations between the two diseases, but also provided genes or gene products that previously have only been investigated separately in the two disease states, and genes or gene products only previously reported to be involved in atherosclerosis. Of the genes or gene products identified, all except two were found to be regulated by LPS. This research illustrates how BMI can help the field of periodontology in the discovery of new directions for

further research and generation of new hypotheses.

References

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25–29.
- bin Ali, A., Zhang, Q., Lim, Y. K., Fang, D., Retnam, L. & Lim, S. K. (2003) Expression of major HDL-associated antioxidant PON-1 is gender dependent and regulated during inflammation. *Free Radical Biology and Medicine* **34**, 824–829.
- Cairo, F., Gaeta, C., Dorigo, W., Oggioni, M. R., Pratesi, C., Pini Prato, G. P. & Pozzi, G. (2004) Periodontal pathogens in atheromatous plaques. A controlled clinical and laboratory trial. *Journal of Periodontal Research* **39**, 442–446.
- Chun, Y. H., Chun, K. R., Olguin, D. & Wang, H. L. (2005) Biological foundation for periodontitis as a potential risk factor for atherosclerosis. *Journal of Periodontal Research* **40**, 87–95.
- Chung, Y. H., Chang, E. J., Kim, S. J., Kim, H. H., Kim, H. M., Lee, S. B. & Ko, J. S. (2006) Lipopolysaccharide from *Prevotella nigrescens* stimulates osteoclastogenesis in cocultures of bone marrow mononuclear cells and primary osteoblasts. *Journal of Periodontal Research* **41**, 288–296.
- Collexis. (2007) Software application available at <http://www.collexis.com>
- Cueto, A., Mesa, F., Bravo, M. & Ocana-Riola, R. (2005) Periodontitis as risk factor for acute myocardial infarction. A case control study of Spanish adults. *Journal of Periodontal Research* **40**, 36–42.
- Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. & Mazo, I. (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* **20**, 604–611.
- Desvarieux, M., Demmer, R. T., Rundek, T., Boden-Albala, B., Jacobs, D. R. Jr., Sacco, R. L. & Papapanou, P. N. (2005) Periodontal microbiota and carotid intima-media thickness: the Oral Infections and Vascular Disease Epidemiology Study (INVEST). *Circulation* **111**, 576–582.
- Dunzendorfer, S., Lee, H. K., Soldau, K. & Tobias, P. S. (2004) Toll-like receptor 4 functions intracellularly in human coronary artery endothelial cells: roles of LBP and sCD14 in mediating LPS responses. *The FASEB Journal* **18**, 1117–1119.
- EntrezGene. (2007) <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene>
- Fiehn, N. E., Larsen, T., Christiansen, N., Holmstrup, P. & Schroeder, T. V. (2005)

- Identification of periodontal pathogens in atherosclerotic vessels. *Journal of Periodontology* **76**, 731–736.
- Fokkema, S. J., Loos, B. G., Hart, A. A. & van der Velden, U. (2003) Long-term effect of full-mouth tooth extraction on the responsiveness of peripheral blood monocytes. *Journal of Clinical Periodontology* **30**, 756–760.
- Garlet, G. P., Martins, W. Jr., Ferreira, B. R., Milanezi, C. M. & Silva, J. S. (2003) Patterns of chemokines and chemokine receptors expression in different forms of human periodontal disease. *Journal of Periodontal Research* **38**, 210–217.
- Gerard, S. & Christopher, B. (1988) Term-weighting approaches in automatic text retrieval. *Information Processing and Management* **24**, 513–523.
- Gibson, F. C. III, Hong, C., Chou, H. H., Yumoto, H., Chen, J., Lien, E., Wong, J. & Genco, C. A. (2004) Innate immune recognition of invasive bacteria accelerates atherosclerosis in apolipoprotein E-deficient mice. *Circulation* **109**, 2801–2806.
- Glenisson, P., Coessens, B., Van Vooren, S., Mathys, J., Moreau, Y. & De Moor, B. (2004) TXTGate: profiling gene groups with text-based information. *Genome Biology* **5**, R43.
- Hajishengallis, G., Sharma, A., Russell, M. W. & Genco, R. J. (2002) Interactions of oral pathogens with toll-like receptors: possible role in atherosclerosis. *Annals of Periodontology* **7**, 72–78.
- Hajishengallis, G., Sojar, H., Genco, R. J. & DeNardin, E. (2004) Intracellular signaling and cytokine induction upon interactions of *Porphyromonas gingivalis* fimbriae with pattern-recognition receptors. *Immunological Investigations* **33**, 157–172.
- Hestvik, A. L., Hmama, Z. & Av-Gay, Y. (2003) Kinome analysis of host response to mycobacterial infection: a novel technique in proteomics. *Infection and Immunity* **71**, 5514–5522.
- HUGO Gene Nomenclature Committee (2007) HGNC (Database). Available at <http://www.gene.ucl.ac.uk/nomenclature/> (accessed 9 October 2007).
- Iwamoto, Y., Nishimura, F., Soga, Y., Takeuchi, K., Kurihara, M., Takashiba, S. & Murayama, Y. (2003) Antimicrobial periodontal treatment decreases serum C-reactive protein, tumor necrosis factor- α , but not adiponectin levels in patients with chronic periodontitis. *Journal of Periodontology* **74**, 1231–1236.
- Janket, S. J., Baird, A. E., Chuang, S. K. & Jones, J. A. (2003) Meta-analysis of periodontal disease and risk of coronary heart disease and stroke. *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontics* **95**, 559–569.
- Jelier, R., Jenster, G., Dorssers, L. C., Wouters, B. J., Hendriksen, P. J., Mons, B., Delwel, R., Kors, J. A., Arias-Vasquez, A., Koudstaal, P. J., Hofman, A., Witteman, J. C., van Duijn, C. M. & Breteler, M. M. (2007) Text-derived concept profiles support assessment of DNA microarray data for acute myeloid leukemia and for androgen receptor stimulation. *BMC Bioinformatics* **8**, 14.
- Kinane, D. F., Riggio, M. P., Walker, K. F., MacKenzie, D. & Shearer, B. (2005) Bacteremia following periodontal procedures. *Journal of Clinical Periodontology* **32**, 708–713.
- Kohane, I. S. (2000) Bioinformatics and clinical informatics: the imperative to collaborate. *Journal of the American Medical Informatics Association* **7**, 512–516.
- Kornman, K. S. & Duff, G. W. (2001) Candidate genes as potential links between periodontal and cardiovascular diseases. *Annals of Periodontology* **6**, 48–57.
- Lalla, E., Lamster, I. B., Hofmann, M. A., Bucciarelli, L., Jerud, A. P., Tucker, S., Lu, Y., Papanoul, P. N. & Schmidt, A. M. (2003) Oral infection with a periodontal pathogen accelerates early atherosclerosis in apolipoprotein E-null mice. *Arteriosclerosis, Thrombosis, and Vascular Biology* **23**, 1405–1411.
- Li, L., Messas, E., Batista, E. L. Jr., Levine, R. A. & Amar, S. (2002) *Porphyromonas gingivalis* infection accelerates the progression of atherosclerosis in a heterozygous apolipoprotein E-deficient murine model. *Circulation* **105**, 861–867.
- Lipscomb, C. E. (2000) Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association* **88**, 265–266.
- Loos, B. G. (2005) Systemic markers of inflammation in periodontitis. *Journal of Periodontology* **76**, 2106–2115.
- Loos, B. G., John, R. P. & Laine, M. L. (2005) Identification of genetic risk factors for periodontitis and possible mechanisms of action. *Journal of Clinical Periodontology* **32** (Suppl. 6), 159–179.
- Luscombe, N. M., Greenbaum, D. & Gerstein, M. (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods of Information in Medicine* **40**, 346–358.
- Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. (2005) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research* **33**, D54–58.
- Mattila, K. J., Pussinen, P. J. & Paju, S. (2005) Dental infections and cardiovascular diseases: a review. *Journal of Periodontology* **76**, 2085–2088.
- MEDLINE (2007) http://www.nlm.nih.gov/databases/databases_medline.html
- Mullenix, P. S., Andersen, C. A. & Starnes, B. W. (2005) Atherosclerosis as inflammation. *Annals of Vascular Surgery* **19**, 130–138.
- van Mulligen, E. M., Diwersy, M., Schmidt, M., Buurman, H. & Mons, B. (2000) Facilitating networks of information. *Proceedings/AMIA Annual Symposium. AMIA Symposium* 868–872.
- Naito, M., Sakai, E., Shi, Y., Ideguchi, H., Shoji, M., Ohara, N., Yamamoto, K. & Nakayama, K. (2006) *Porphyromonas gingivalis*-induced platelet aggregation in plasma depends on Hgp44 adhesin but not Rgp proteinase. *Molecular Microbiology* **59**, 152–167.
- Ng, C. J., Bourquard, N., Grijalva, V., Hama, S., Shih, D. M., Navab, M., Fogelman, A. M., Lusis, A. J., Young, S. & Reddy, S. T. (2006) Paraonase-2 deficiency aggravates atherosclerosis in mice despite lower apolipoprotein-B-containing lipoproteins: anti-atherogenic role for paraonase-2. *The Journal of Biological Chemistry* **281**, 29491–29500.
- Ng, C. J., Shih, D. M., Hama, S. Y., Villa, N., Navab, M. & Reddy, S. T. (2005) The paraonase gene family and atherosclerosis. *Free Radical Biology and Medicine* **38**, 153–163.
- Nichols, T. C., Fischer, T. H., Deliyargyris, E. N. & Baldwin, A. S. Jr. (2001) Role of nuclear factor-kappa B (NF-kappa B) in inflammation, periodontitis, and atherogenesis. *Annals of Periodontology* **6**, 20–29.
- Nikitin, A., Egorov, S., Daraselia, N. & Mazo, I. (2003) Pathway studio – the analysis and navigation of molecular networks. *Bioinformatics* **19**, 2155–2157.
- NLM (2007) <http://www.nlm.nih.gov/>
- Nordlie, M. A., Wold, L. E. & Kloner, R. A. (2005) Genetic contributors toward increased risk for ischemic heart disease. *Journal of Molecular and Cellular Cardiology* **39**, 667–679.
- Novichkova, S., Egorov, S. & Daraselia, N. (2003) Medscan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics* **19**, 1699–1706.
- Ott, S. J., El Mokhtari, N. E., Musfeldt, M., Hellmig, S., Freitag, S., Rehman, A., Kuhbacher, T., Nikolaus, S., Namsolleck, P., Blaut, M., Hampe, J., Sahly, H., Reinecke, A., Haake, N., Gunther, R., Kruger, D., Lins, M., Herrmann, G., Folsch, U. R., Simon, R. & Schreiber, S. (2006) Detection of diverse bacterial signatures in atherosclerotic lesions of patients with coronary heart disease. *Circulation* **113**, 929–937.
- Ozer, E. A., Pezzulo, A., Shih, D. M., Chun, C., Furlong, C., Lusis, A. J., Greenberg, E. P. & Zabner, J. (2005) Human and murine paraonase 1 are host modulators of *Pseudomonas aeruginosa* quorum-sensing. *FEMS Microbiology Letters* **253**, 29–37.
- PathwayStudio. (2007) <http://www.ariadnegenomics.com/products/pathway.html>
- Peake, P. W., Shen, Y., Campbell, L. V. & Charlesworth, J. A. (2006) Human adiponectin binds to bacterial lipopolysaccharide. *Biochemical and Biophysical Research Communications* **341**, 108–115.
- Pihlstrom, B. L., Michalowicz, B. S. & Johnson, N. W. (2005) Periodontal diseases. *Lancet* **366**, 1809–1820.
- PubMed (2007) <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>
- Ren, L., Jin, L. & Leung, W. K. (2004) Local expression of lipopolysaccharide-binding protein in human gingival tissues. *Journal of Periodontal Research* **39**, 242–248.
- van Rijn, M. J., Bos, M. J., Yazdanpanah, M., Isaacs, A., Arias-Vasquez, A., Koudstaal, P. J., Hofman, A., Witteman, J. C., van Duijn, C. M. & Breteler, M. M. (2006) Alpha-adducin polymorphism, atherosclerosis, and

- cardiovascular and cerebrovascular risk. *Stroke* **37**, 2930–2934.
- Rodriguez Esparragon, F., Hernandez Trujillo, Y., Macias Reyes, A., Hernandez Ortega, E., Medina, A. & Rodriguez Perez, J. C. (2006) Concerning the significance of Paraoxonase-1 and SR-B1 genes in atherosclerosis. *Revista Española de Cardiología* **59**, 154–164.
- Rosner, D., Stoneman, V., Littlewood, T., McCarthy, N., Figg, N., Wang, Y., Tellides, G. & Bennett, M. (2006) Interferon-gamma induces Fas trafficking and sensitization to apoptosis in vascular smooth muscle cells via a PI3K- and Akt-dependent mechanism. *The American Journal of Pathology* **168**, 2054–2063.
- Ross, R. (1999) Atherosclerosis – an inflammatory disease. *New England Journal of Medicine* **340**, 115–126.
- Rothenbacher, D., Muller-Scholze, S., Herder, C., Koenig, W. & Kolb, H. (2006) Differential expression of chemokines, risk of stable coronary heart disease, and correlation with established cardiovascular risk markers. *Arteriosclerosis, Thrombosis, and Vascular Biology* **26**, 194–199.
- Salton, R. (1989) *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston: Addison-Wesley Longman Publishing Co Inc.
- Schoppet, M., Sattler, A. M., Schaefer, J. R. & Hofbauer, L. C. (2006) Osteoprotegerin (OPG) and tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) levels in atherosclerosis. *Atherosclerosis* **184**, 446–447.
- Schuemie, M., Chichester, C., Lisacek, F., Coute, Y., Roes, P. J., Sanchez, J. C., Kors, J. & Mons, B. (2007) Assignment of protein function and discovery of novel nucleolar proteins based on automatic analysis of MEDLINE. *Proteomics* **7**, 921–931.
- Shatkay, H. (2005) Hairpins in bookstacks: information retrieval from biomedical text. *Brief Bioinform* **6**, 222–238.
- Sheikine, Y., Bang, C. S., Nilsson, L., Samnegard, A., Hamsten, A., Jonasson, L., Eriksson, P. & Sirsjo, A. (2005) Decreased plasma CXCL16/SR-PSOX concentration is associated with coronary artery disease. *Atherosclerosis* **188**, 462–466.
- Silva, T. A., Garlet, G. P., Lara, V. S., Martins, W., Jr., Silva, J. S. & Cunha, F. Q. (2005) Differential expression of chemokines and chemokine receptors in inflammatory periapical diseases. *Oral Microbiology and Immunology* **20**, 310–316.
- Smalheiser, N. R. & Swanson, D. R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine* **57**, 149–153.
- Soder, P. O., Soder, B., Nowak, J. & Jogestrand, T. (2005) Early carotid atherosclerosis in subjects with periodontal diseases. *Stroke* **36**, 1195–1200.
- Spahr, A., Klein, E., Khuseynova, N., Boeckh, C., Muehe, R., Kunze, M., Rothenbacher, D., Pezeshki, G., Hoffmeister, A. & Koenig, W. (2006) Periodontal infections and coronary heart disease: role of periodontal bacteria and importance of total pathogen burden in the Coronary Event and Periodontal Disease (CORODONT) study. *Archives of Internal Medicine* **166**, 554–559.
- Tabrizi, F., Buhlin, K., Gustafsson, A. & Klinge, B. (2007) Oral health of monozygotic twins with and without coronary heart disease: a pilot study. *Journal of Clinical Periodontology* **34**, 220–225.
- Taubman, M. A., Valverde, P., Han, X. & Kawai, T. (2005) Immune response: the key to bone resorption in periodontal disease. *Journal of Periodontology* **76**, 2033–2041.
- Tripi, L. M., Manzi, S., Chen, Q., Kenney, M., Shaw, P., Kao, A., Bontempo, F., Kammerer, C. & Kamboh, M. I. (2006) Relationship of serum paraoxonase 1 activity and paraoxonase 1 genotype to risk of systemic lupus erythematosus. *Arthritis and Rheumatism* **54**, 1928–1939.
- Veillard, N. R., Steffens, S., Pelli, G., Lu, B., Kwak, B. R., Gerard, C., Charo, I. F. & Mach, F. (2005) Differential influence of chemokine receptors CCR2 and CXCR3 in development of atherosclerosis in vivo. *Circulation* **112**, 870–878.
- Vohra, R. S., Murphy, J. E., Walker, J. H., Ponnambalam, S. & Homer-Vanniasinkam, S. (2006) Atherosclerosis and the lectin-like oxidized low-density lipoprotein scavenger receptor. *Trends in Cardiovascular Medicine* **16**, 60–64.
- Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W. & Povey, S. (2002) Guidelines for human gene nomenclature. *Genomics* **79**, 464–470.
- Weeber, M., Vos, R., Klein, H., De Jong-Van Den Berg, L. T., Aronson, A. R. & Molema, G. (2003) Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association* **10**, 252–259.
- Xu, H., Xu, W., Chu, Y., Gong, Y., Jiang, Z. & Xiong, S. (2005) Involvement of up-regulated CXC chemokine ligand 16/scavenger receptor that binds phosphatidylserine and oxidized lipoprotein in endotoxin-induced lethal liver injury via regulation of T-cell recruitment and adhesion. *Infection and Immunity* **73**, 4007–4016.
- Yoshie, H., Kobayashi, T., Tai, H. & Galicia, J. C. (2007) The role of genetic polymorphisms in periodontitis. *Periodontol 2000* **43**, 102–132.

Address:
 Kristina Hettne
 Department of Medical Informatics
 Erasmus University Medical Center Rotterdam
 PO Box 2040, 3000 CA
 Rotterdam
 The Netherlands
 E-mail: k.hettne@erasmusmc.nl

Clinical Relevance

Scientific rationale for the study: New tools from the BMI field may help to generate new hypotheses on how periodontitis and cardiovascular diseases are linked.

Principal findings: We identified 16 genes possibly to play a role in both

diseases, with PON1 as the primary candidate. Genetic variations in PON1 have been associated with inflammatory conditions and atherosclerosis. Therefore, PON1 and other genes identified in this study may open up new research possibilities.

Practical implications: A combined approach of association-based and natural language processing-based literature mining may help to understand the link between periodontitis and cardiovascular diseases.