A More Intelligent Literature Search
- Shoulders of Giants

Michael G King, Alison Van Bree
Sunshine Health and Dental Superclinic
429 Ballarat Rd, Sunshine,
Victoria, Australia 3020.

***Abstract*—Although the topic of study relates to an environmental/health issue, it is the methodology described which serves to showcase an embryonic form of a new "more intelligent" protocol of search algorithm. Through the implementation of this algorithm, an extensive automated literature base yielded a single credible solution to a previously unsolved problem. Faced with a distressing but entirely unexplained incidence of birth defects, the proposed model of knowledge scavenging worked through acknowledged gaps in understanding of increased (phosphate) fertilizer, enabled the template of known facts regarding the interactions of phosphates with the processes of mammal (and other animal) growth, of metabolic function, and of neurological development, and delivered a causal model which would not, at least not easily, derive from current literature search methods. Illustrating the practical value of a step forwards in the design of intelligent literature search, the present study provides a candidate cause to explain a cluster of bovine deformity.**

Keywords—automated literature search; database; search algorithm; Craniosynostosis; fibroblast growth factor receptor

I. INTRODUCTION

The levels of animal production from these pastures are impressive. Yet we are not quite sure just why this happens. . . . . There are gaps in our understanding of soil/animal/plant relationships, particularly when soil fertility is increased and grazing pressure is intensified [1].

It is easy to acknowledge that there are "gaps in our understanding" of any, indeed of every, topic in the bio-related disciplines, however a more accurate description of "what is not presently known" might be that there is likely the totality of the required knowledge already discovered but located in a fragmented array of publications located in a variety of nominally discrete disciplines. Such an assertion acknowledges the entirely vast number of bio-medical-chemical publications being added to by millions of further publications each year (in the date range 2012 – 2013 the number of hits returned for PubMed data base is 2056910). Thus it would be arrogance, hubris, or at least out of step with Occam's razor to assume that whatever is being investigated actually includes "gaps" in total human, published knowledge. It is the more likely case that somewhere there is the missing information but it is not findable by conventional search. The present paper looks to this problem, introduces a strategy of solution, and provides a precise example illustrating the benefits of a more intelligent literature search.

Animal Birth Defects: Acorn Disease.

Recent examples of clusters of birth defects among animals (from 1% to 50% incidence in certain areas in Australia during the recent drought) had no obvious explanation. The defects listed include facial deformities and limb malformation. Genetic issues were investigated and eliminated. In short, the animal scientists remain baffled. Conventional investigative approaches might be summarized by the following steps:

a) investigating known information about animal defects, and

b) considering the information even more fully, in ever increasing detail.

This approach makes intuitive sense, however in the situation where there no likely candidate explanation of the problem, and when none seems to be emerging from the investigation, it makes sense to look outside the square. For the purposes of the present study "outside the square" might mean looking at human birth defects. It makes an interesting, even ironical, twist of laboratory processes to use investigations on humans to help solve animal disease. Even further "outside the conceptual square" would be an entirely different approach to the computer process of literature search – a methodology designed with the biochemical/medical sciences in mind.

Human Deformities studies

The study of animal birth defects differs from that of human deformities: the former is largely of technical interest, while the latter has an emotive overlay; the former is a bump in an established (or ignored, or even suppressed) statistical data stream, while the latter is something which is disturbingly abhorrent at any frequency. Given the affective incentive, research into causes and foci of action of candidate causative agents is likely to be more vigorous and extensive in the case of human birth defects. Therefore in considering the situation of a putative increase in bovine birth defects, it is appropriate to cast the knowledge net wider and trawl for answers in the human-based data.

The calves born under the currently-studied cluster of "acorn disease" features deficient limb development, as well as cranio-facial deformity. Corresponding to these anomalies deficient limb development in humans is included in studies of the broad range of conditions grouped under the heading of pre-natal rickets, while an understanding of the cranio-facial defects in the target calves may benefit from a review of Craniosynostosis in humans. This review will take the two above-named domains of study in reverse order.

Craniosynostosis Studies

There are two putative forms of Craniosynostosis, both forms being conventionally linked to genetic mutations:

The Saethre-Chotzen and the Muenke craniosynostoses,. . . . Patients with Saethre-Chotzen and Muenke syndromes carry mutations in the TWIST and fibroblast growth factor receptor (FGFR) 3 genes, respectively [2].

And pursuing the mutant Growth Factor (FGF, FGFR etc) variant of this disorder, it is noted that:

Mutations in the FGFR gene family have been linked to a series of syndromes (the craniosynostosis syndromes) whose primary phenotype involves aberrant development of the craniofacial skeleton. . . . Unfortunately, studies attempting to link expression of mutant FGFRs with changes in cellular phenotype have yielded conflicting results [3].

At first glance, human Craniosynostosis studies would appear to have little potential to cast light upon the current acorn-diseased calves: first the human studies are "conflicting" and secondly, genetic factors in the diseased calves have been all but ruled out by veterinary review (private communication from investigating group). However there are many physiological and biochemical steps between genetic information and limb or facial formation, and when the causal emphasis upon mutant genes is put to one side, there remains a credible common thread: whether genetic or not, perturbations are found in the Growth Factor component of the complex process of bone development. Observe that the following quotation reads equally well whether or not genetic mutation is implied: Craniosynostosis syndrome-linked FGFR mutations have been shown to be gain of function in terms of receptor activation and have been presumed to result in increased levels of FGF/FGFR signalling [3].

Although Hatch's [3] specific proposition is that "increased levels of FGF/FGFR signalling" is the cause of Craniosynostosis, it would be sensible to keep the net wider and work under the more general rubric of "Growth Factor(s)", naming neither "fibroblast Growth Factor" (FGF) nor the "Receptor" (FGFR) as the specific focus of attention.

In summary, to this point, it appears relevant to search the human literature for similar or parallel defects in the quest for an understanding of bovine birth defects (acorn disease). Furthermore, it follows from an open-minded reading of the Craniosynostosis literature that the search could credibly consider links to Growth Factor (including specific chemical effects thereon), and also other studies of normal or abnormal limb and cranio development.

In seeking this "chemical cause", it was the underlying a priori position, that the chance of a "rare" causative agent in the sense of truly unusual or outstandingly complex was all but eliminated by definition. If there were a rare agent at work, its very rareness would bring it to the forefront of likely candidate causes - its rareness would be its downfall. Given the apparently mysterious nature of clusters of craniosynostosis, accepting that an answer has never immediately come to hand then the likely candidate cause logically should be some agent which will be masquerading as normality. And then following this position, the implicit "normality" of the causal factor leads to the expectation that the topic is well covered in the literature, albeit without the knowledge or recognition of the authors (and hence the literature search systems).

II. METHOD

The entire biochemical/biomedical data bases contain information from disparate sub-disciplines (branches of medicine and branches of the biochemical sciences). This knowledge base which well exceeds 100,000 items is credibly regarded as the work of *Giants*. The bio-medical data base is approached for specific "problems". These problems require an "answer", where the notion of an "answer" is conceptually regarded as a chemical agent which is either a causal vector, or an intervention. For the Shoulders of Giants (SOG) project it is the assumption that the answer is mentioned somewhere in this total literature base, along with oblique cues which do not directly link this agent to the notional "problem".

The hypothesis of the SOG algorithm is that the target information (a "candidate answer") may, at any stage of a literature search, be only one step away. The problem for the investigator is that, by definition, the direction of that step is neither known nor knowable. If the direction were known, then the investigator would immediately guide a conventional search in that direction.

The "conventional" computer based search is set to trawl data bases for instances of a target word or phrase. Thus, on the bio-medical data bases the word SMOKING occurs among perhaps 25,000 abstracts. The search is refined, in the end, by pairing the target with another limiter (for example ASTHMA, itself having, say, 16,000 hits) thus SMOKING and ASTHMA occurs only in only 2688 abstracts – still too many to review. Further limitations will bring this number down to a manageable size (notionally around 100 abstracts) which is hopefully highly focused (example SMOKING, ASTHMA and CAFFEINE might feature together in a human-perusable set of abstracts rather less than 100 in number).

Modern search strategies assist the search for parallel studies, by locating matches with the key words of other papers in the data base. These assisted searches are not designed to find an answer to a problem when the "answer" is as-yet unknown (unknown in the sense of not been linked to the key words of the "question" underlying the seeding search).

The following observations derive from a word-by-word review of around 30,000 abstracts from the medical science. The process employed was a dedicated but flexible computer program designed precisely for application in the medical/biological sciences where "causal links" are sought from an array of knowledge too large for normal contemplation. In this instance the program seeks to locate common factors linked to the human defect Craniosynostosis this factor being shared with other birth defects in animals. A shared putative causal factor would tend to support the notion of environmental cause.

The (SOG) project has devised a search strategy, or "algorithm", which can suggest answers even though these offerings have not been previously linked to the question – at

least not linked by bridges which are readily discovered through normal search processes.

The SOG project comprises a series of computer programs designed to:

a) Interact with the entire bio-medical data bases (BDB) and extract multiple large sub-sets of literature abstracts

b) isolate these large subsets of the BDB (perhaps thousands of abstracts each) for closer scrutiny

c) take the results of (b) and compare two or more notionally disparate subsets of BDB for overlap

d) provide a manageable few candidate "answers" (really "directions" of likely fruitful investigation) which can be followed by conventional investigative processes.

e) *At the current phase, collect together and print in RTF format (from the "extracted" data sets - step "a", above) sets of abstracts which feature a target "term" (usually a single word or phrase). This step is, of course, available from conventional data search (for example "Smoking" papers which mention "Asthma")*

Step "c" – comparison of two data sets.

This step is at the heart of the SOG program suite. Ultimately pairing for meaningful matches is the core process to deal with the potentially explosive output derived from conducting multiple, nominally unrelated, literature searches.

Once two data sets of any size (eg 20,000 abstracts) have been nominated (at present by human-initiated keyboard command (but in the future by an algorithm itself generated from the initial "problem") then the program can remove a pre-determined list of "trivial" terms, and review the two filtered data sets for meaningful overlap.

III. Results

First common Term.Review of the first cut of the literature (around 30,000 abstracts), points to a simple dichotomy: either the answer remains unknown, or the causative vector contains the term "phosphate". When all "trivial" terms were eliminated from the abstracts reviewed, this term was the only common word which linked the abstracts which derived from the various domains.

The literature review tendered the term "phosphate" as the only common word linking the various abstract groups. Taking the positive view that this link is "non-trivial" and therefore worth considering, the probable cause is some level of organo-phosphates ingested (exposed?) pre-partum. The relevant literature could now be searched in a more conventional manner using standard abstract-sifting strategies (eg "phosphate" with "craniosynostosis") or the down-loaded gamut of abstracts could be further trawled using the above-mentioned abstract-reviewing program.

Use of First Common Term A new and more focused review of the literature confirmed the non-controversial direct effects of various organophosphates, and the more indirect effects of putative metabolites of organophosphates (including the established potentially deleterious effects of deviant levels of "normal" phosphate metabolites) are noted in the literature.

A future review could, by currently-available search methodologies, collect and display all relevant references relating to:

a) the overlap of animal defects and craniosynostosis pointing to "phosphate" as the common factor;

b) the established links between FGF and craniosynostosis, including the relatively weak genetic explanation for FGF-related forms of the disorder: Mutations in the FGFR gene family have been linked to a series of syndromes (the craniosynostosis syndromes) [3].

c) *the established links between organophosphates and altered FGF activity.*

IV. Discussion

The putative organophosphate-linked explanation of craniosynostosis raises questions about the delivery vector, and the identity of the putative phosphate. It is relevant that the application of extra phosphate has been lauded as appropriate during drought years, with the goal of enhancing otherwise impoverished crops. It is already noted in the pro-phosphate agri-literature that the precise mechanisms of this method are not clearly understood – and that "gap in understanding" referred only to the crop-producing qualities. Little or no consideration appeared to be accorded to effects of additional phosphates outside the domain of grass growth.

The current explanation (that increased phosphate is a causal factor in birth defects) has the advantage of matching at least one of the conceptual conditions precursive to the present study: it was anticipated that a genuinely novel chemical compound would NOT have been the cause of the defects, and more particularly, that a popular and presumed safe compound must be the culprit.

V. Conclusion

Noting that the final validity of the science under scrutiny is not the central issue, but merely an illustration of the potential results from a step forward in intelligent search algorithms, it is presently submitted that a credible *a priori* case for organophosphate activity as the causal factor of both craniosynostosis and animal birth defects exists. This proposition is based upon non-controversial and established literature however the proposed "answer" would not be arrived at by conventional literature search alone. The underlying search strategy firstly accepts that an entirely huge and unwieldy quantity of scientific investigation has been carried in an ever widening and ever more isolated myriad of separate fields.

It is the primary and founding proposition for the presently developed approach that almost any "advance" in understanding, even in the development of treatments and cures, will be seen (with the benefit of hindsight) to have touched on and brought together ideas which have been described in the literature – but these separate components have not been linked together. That is to say that many future (especially bioehcmical) developments will have been hiding in plain sight.

For the present first step (illustrating that a tentative solution can be found for a problem which has defied solution by

conventional focused research by experts), the elegance of the uncovered "solution" is firstly that a mundane cause for presumed genetic conditions has not previously been signaled nor sought, and (reversing the vectorial sense of the research) the studies of phosphate and limb or other development had not been interpreted as candidate explanations for putative genetic conditions. The reason why this answer has not been envisaged by others is because of the (likely) assumption that a mutagenic agent rather than an FGF-interactive agent was sought. The SOG algorithm makes no assumptions but merely seeks links where none were previously imagined; it is then up to the human scientist to consider the credibility and ultimately the validity of the suggested links.

In summary the SOG-derived propositions were:

- Cranial malformation at birth can be due to fibroblast (fibroblast growth factor) aberrations.
- Phosphates are involved in aberrant pre-natal development, possibly due to an interaction with fibroblast growth factor. Most common studies relate to low phosphate levels and bone deformities (hypophosphate conditions) however high serum phosphate levels are equally linked to relevant deleterious outcomes [4,5] of hyperphosphatemia and impaired skeletogenesis.
- Broader ranges of experiments beyond the neo-natal stage show a variety of interactions between phosphates (typically organophosphates) and various fibroblast or fibroblast growth factor activity.
- Prenatal exposure to aberrant phosphate conditions can lead to a fibroblast-mediated condition of rickets.

The strength of the present proposition lies in that it is based upon non-controversial science. In addition, there is not an obvious competing explanation for the coincidence of these disorders.

While it may be of interest to have a putative explanation of a mysterious cluster of animal or human birth defects, the true strength of the present study is that it showcases a novel method of investigation that was precisely designed to serve the bio-medical domain. The validity of the analytical procedure depends entirely upon the proposition that somewhere, in two or more nominally disparate domains, there already co-exist studies which can be brought together to cast light upon currently mysterious questions.

The illustrated "solution" (whether ultimately accepted or rejected) starkly shows a solution which would be invisible to current literature search algorithms which are not based upon the approach of comparing two or more disparate data sets in order to find the "hidden, but existing" solution to problems.

Future work would ideally build the "next step" into the automated search process. The process as applied to the present problem analysed the entire (that is all that were available as abstracts) literature base pertaining to a single search term, and then proceeded to list the frequency of any "non-trivial" words/phrases. This resulting list of over 4000 items was matched with another list of over 11,000 items and the resulting overlap of less than 300 items was compared by human consideration and the result was just one potential explanatory term. The next phase of the SOG project would be to have the computer program "learn" more about filtering non-explanatory terms. However the final strategy which is at this stage beyond the resources of the authors would be to enter a topic of interest (effectively a single search term) and the SOG suite automatically access the data bases, extract the likely immense set of literature, peruse these for terms worth pursuing, and follow with as many "matching search result sets" as necessary. The principal author would be pleased to cooperate in whatever manner may lead to the fruition of this new concept, while, returning to what can be learned from the present illustration of part of this whole strategy effectively done "by hand": in many cases, the truth is already out there, but in fragmented forms in diverse places.

REFERENCES

[1] Sale, P. *The Pasture Productivity Revolution*. Department of Agricultural Sciences. LaTrobe University, Melbourne, Australia. 12pp. 2007.

[2] Jadico SK. Huebner A. McDonald-McGinn DM. Zackai EH. Young TL. Ocular phenotype correlations in patients with TWIST versus FGFR3 genetic mutations. *Journal of AAPOS: American Association for Pediatric Ophthalmology & Strabismus.* 10(5):435-44, 2006 Oct.

[3] Hatch NE. Hudson M. Seto ML. Cunningham ML. Bothwell M. Intracellular retention, degradation, and signaling of glycosylation-deficient FGFR2 and craniosynostosis syndrome-associated FGFR2C278F. *Journal of Biological Chemistry*. 281(37):27292-305, 2006 Sep.

[4] Garringer HJ. Fisher C. Larsson TE. Davis SI. Koller DL. Cullen MJ. Draman MS. Conlon N. Jain A. Fedarko NS. Dasgupta B. White KE. The role of mutant UDP-N-acetyl-alpha-D-galactosamine-polypeptide N-acetylgalactosaminyltransferase 3 in regulating serum intact fibroblast growth factor 23 and matrix extracellular phosphoglycoprotein in heritable tumoral calcinosis. *Journal of Clinical Endocrinology & Metabolism*. 91(10):4037-42, 2006 Oct.

[5] Sitara D. Razzaque MS. Hesse M. Yoganathan S. Taguchi T. Erben RG. Juppner H. Lanske B. Homozygous ablation of fibroblast growth factor-23 results in hyperphosphatemia and impaired skeletogenesis, and reverses hypophosphatemia in Phex-deficient mice. *Matrix Biology*. 23(7):421-32, 2004 Nov.